2013

Novel computational tools and utilization of the gut microbiota for phylogeographic inference

Sarah Michelle Hird
Louisiana State University and Agricultural and Mechanical College, shird1@tigers.lsu.edu

NOVEL COMPUTATIONAL TOOLS AND UTILIZATION OF THE GUT MICROBIOTA FOR PHYLOGEOGRAPHIC INFERENCE

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Biological Sciences

by
Sarah Michelle Hird
B.S., University of Idaho, 2005
M.S., University of Idaho, 2008
May 2013

Acknowledgments

This document is years in the making and the effort of many people. First, my advisors, Dr. Robb Brumfield and Dr. Bryan Carstens, have my eternal gratitude for being the perfect yin and yang. Together, they have allowed me the space and the structure to work at my own pace yet push my boundaries. They were always there when I needed guidance of any kind. Bryan, in particular, first encouraged me to learn Perl and I can't express how grateful I am that he did. Robb invested in me as a student and thus granted me access to the LSU Museum of Natural Science (LSUMNS). I'd also like to thank my Master's advisor, Dr. Jack Sullivan, for teaching me good habits and starting me on this path in the first place. The NSF and LSU funded my degree and research and I very much appreciate that.

My committee members, Dr. Brent Christner, Dr. Brygg Ullmer and Dr. Robert Rohli have all been available and informative whenever I asked. In particular, Dr. Christner introduced me to the world of microbial ecology and discussions with Dr. Ullmer helped solidify my interest in bioinformatics. The LSUMNS faculty, staff and students are a group of unbelievably amazing naturalists and scientists; they are directly responsible for the (comparatively very small) amount of natural history I know. The access to and support from the Bird Lunch group was integral in guiding my reasoning and justification for the bird gut chapters; a special thanks to the extremely helpful comments of Dr. J. Van Remsen and Dr. Fred Sheldon.

I have been fortunate to work with many wonderful collaborators and post-doctoral researchers. Dr. Maggie Hanes (Koopman) was a generous friend, mentor and collaborator; I strive to be like her in more ways than one. Dr. John McCormack included me in many projects that were of great interest and value to me. He and Dr. Amanda Zellmer offered friendship, intellect and perspective on many issues. Dr. Erica Tsai provided a lot of useful advice regarding pursuing bioinformatics and thoughtfully answered any questions I had. Many other professionals had a strong positive influence on me and my education/research; thank you Dr. Jeremy Brown, Steve Cardiff, Donna Dittmann, Dr. Brian Smith, Dr. Brad Chapman and Daniel Ence.

The chapters on bird gut microbiota would never have happened without the enthusiasm, encouragement and hard work of many museum graduate students. Cesar Sanchez conducted much of the fieldwork and contributed both physically and intellectually to the Neotropical bird project. Michael Harvey, Dr. Gustavo Bravo and all the other field assistants who collected many intestinal tracts for me and I appreciate their hard work more than I can write. Graduate students at the museum and in the Department of Biological Sciences contributed much to my education. I gratefully acknowledge the following individuals for friendship on top of help with reading groups, paper edits, talk critiques, good/bad ideas, etc., especially over coffee or cocktails: Melissa DeBiasse, Eric Rittmeyer, Verity Mathis, Lorelei Patrick, Cathy Newman, Clare Brown, Vivien Chua.

All past and present members of both the Brumfield and Carstens labs provided an amazing work environment and helped me professionally by reading drafts of manuscripts, discussing ideas, critiquing talks, and much more. Thank you – Noah Reid, John McVay, Tara Pelletier, Jordan Satler, Dr. James Maley, Andres Cuervo, Caroline Duffie, Dr. Gustavo Bravo, Michael Harvey, Glenn Seeholzer, John Mittermeier. These people are also wonderful human beings and good friends.

Emotional support and well-being can be attributed to my many good friends at LSU and beyond, especially Tara Pelletier, John McVay and Megan McVay. Nala is the four-legged definition of stress relief.

My family is amazing and everything that I am is because of them. Joan Hopkins, my mother, is a paragon of efficiency and determination and a role model that I am grateful for every day. She and my sisters, Katie Hird and Stephanie Thompson, comprise my heart.

Finally, none of this would exist without Noah Mattoon Reid, who makes everything in my life, including this dissertation, better. He is my best friend, my husband, my mentor, my sounding board, my advocate, my editor, my support and my sanity. Thank you, thank you, thank you.

Table Of Contents

Abstract

Genetic data are frequently responsible for biological insight and recent advances in sequencing technology (high-throughput sequencing; HTS) have created massive DNA-sequence based datasets. While these technologies are invaluable, there are many analytical and application issues that need to be addressed. With these data we can ask and answer novel biological questions that were previously inaccessible.

One major challenge in applying HTS to biological questions is data management: the file formats and sizes are foreign to many primary researchers. In the second and third chapters of this dissertation, I introduce two pieces of software that allow researchers to utilize HTS with minimal time investment. PRGMATIC (Chapter 2) is a pipeline that collates raw HTS data into a more traditional and useable format: two diploid alleles for a given locus. LOCINGS (Chapter 3) uses these loci, the alignments from which the loci were called and demographic data to display and output important summary statistics. This program also reformats appropriate loci into three widely used biological file formats.

Chapters 4 and 5 focus on a novel application of HTS to phylogeographic inference. The collective set of microbial organisms on and inside vertebrates (the microbiota) is a vast genetic resource that is poorly understood. What factors shape these communities? Chapter 4 uses an avian brood parasite (Brown-Headed Cowbird) to naturally decouple parental genetics and early environment. Cowbird gut microbiota do not cluster with each other in multivariate space. They also do not strongly affiliate with host species. Age and sampling locality are most strongly associated with the gut microbiota. Chapter 5 looks for host taxonomic and spatial signals in a more broadly sampled dataset of 60 species sampled across Costa Rica. Here, host taxonomy is most significantly associated with gut microbiota and ecological variables like host diet and foraging strata are secondarily important.

Together, these chapters present novel tools and uses of HTS for evolutionary inference. The two programs, PRGMATIC and LOCINGS, allow primary researchers to utilize HTS easily. The two bird datasets, cowbirds and Costa Rican birds, demonstrate how analyzing the microbiota with HTS can provide and address novel evolutionary questions.

Chapter 1.
Introduction: High-Throughput Sequencing, Computational Tools and Host-Associated Microbiota*

Biologists utilize genetic data to make inferences about evolutionary history (phylogeny), physiology, disease/health, demography and many other processes. DNA sequences can tell us about past events, like migration, population bottlenecks and genetic drift, and they can tell us about the relationships among organisms at various taxonomic levels and time scales. The importance of these data cannot be overstated and the introduction of high-throughput sequencing (HTS) technologies, starting in early 2005, has revolutionized many aspects of biological research. Through various molecular, technical and computational advances, HTS platforms have increased the amount of sequencing data produced by a single use of a machine by over seven orders of magnitude while decreasing the cost per base by six orders of magnitude (Glenn, 2011). This massive change in the amount of data biological researchers can obtain has led to exciting new questions and applications (Altschul et al., 1990) as well as new challenges (McPherson, 2009). The aim of this dissertation is to collect empirical HTS datasets and develop computational tools to answer questions across evolutionary biology (Fig. 1.1).

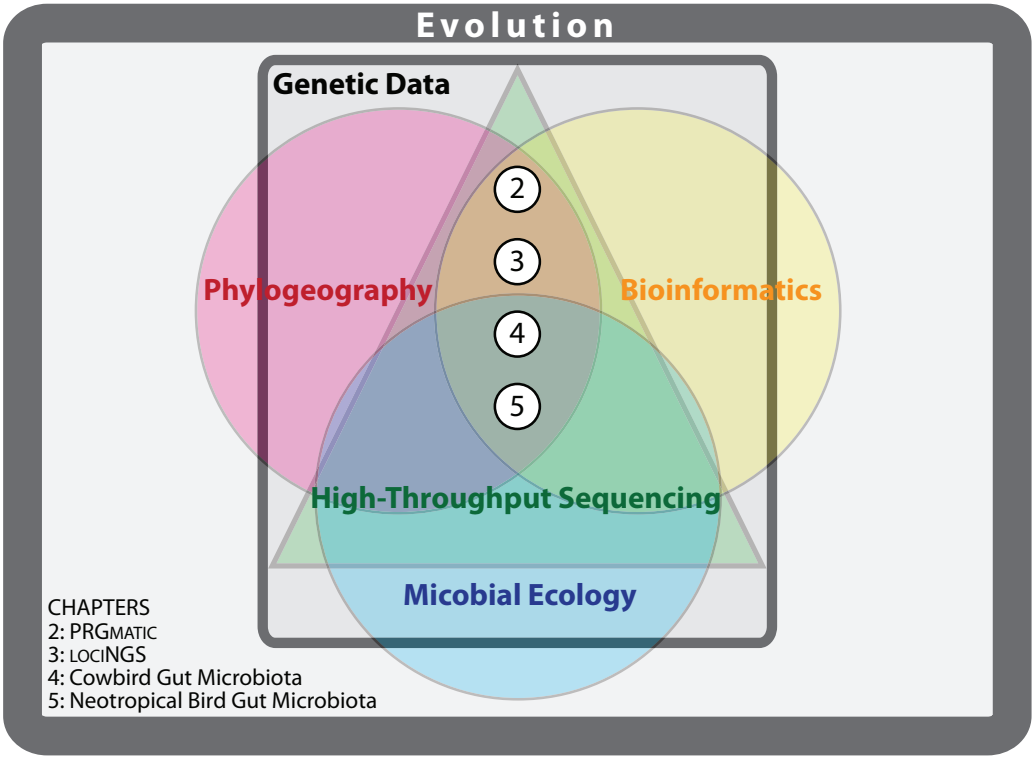


Figure 1.1. Venn diagram of my research interests and where the four research chapters of this dissertation fall within them.

* Portions of this chapter previously appeared as: McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generations sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution*, 66(2): 526-538. Reproduced with publisher's permission.

1.1. HIGH-THROUGHPUT SEQUENCING

Simply storing unprocessed HTS data requires significant computer memory and often hardware upgrades or remote (i.e., online or ‘‘cloud’’) storage. Whereas a reasonably large Sanger dataset may contain 500 sequences, typical 454 and Illumina runs generate 1×10^6 to 2×10^9 sequences, and these numbers are increasing rapidly as the sequencing platforms are refined. Data set sizes now are measured in terabytes and file transfer is often conducted through normal postal mail due to the uploading/downloading time and cost of sending files through the internet or cloud. However, logistical difficulties aside, these numbers are in some ways deceptively large, because another major difference between HTS data and Sanger sequence data is the quality of the reads (i.e., many low quality reads are discarded).

Chromatograms are an intuitive way to assess the quality of a given Sanger sequence because the colored peaks are a reflection of the strength of each nucleotide's signal. Frequently, with Sanger data, a human being has evaluated all or most of the bases called by the sequencer. This is not possible with even a small HTS dataset. HTS quality scores are an integral part of the sequence data itself, and come either as a series of integers or letters corresponding to every base called by the HTS platform. These are very different paradigms: Sanger data provides, in essence, a more or less true snapshot of a pool of amplicons at every position. HTS data provides a slightly imperfect representation of some of the amplicons from a sample, along with their associated quality scores. This is why coverage (i.e., the number of reads that support a specific base call) is critically important with HTS data. Coverage affords confidence that every DNA fragment in a pool will be sequenced – and enough times to determine heterozygous positions. Coverage also ensures that HTS sequencing error can be detected and distinguished from true biological heterozygosity.

There are more differences between Sanger sequencing and HTS than just output quantity and the importance of coverage. Biological questions must be posed such that short reads or single nucleotide polymorphisms (SNPs) can address them and funding is spent in larger units than that for Sanger sequencing. Additionally, new tools must be developed and used that are capable of handling the massive amounts of data and new file formats, for example, the Sequence Alignment/Map format (Li et al., 2009). We also must explore the ways these new tools allow us to investigate the world around us.

1.2. COMPUTATIONAL TOOLS

Although the rate at which we can gather genetic data has increased rapidly in recent years, the ability to analyze these data has lagged behind. This “next-generation gap” between data acquisition and data analysis is a limiting factor in many cases (McPherson, 2009); additional money (to purchase commercial software or computational help) and/or a bioinformatician are frequently necessary. Before any biological information can be gleaned from the data, they must be made suitable for analysis. This process includes quality checking each sequence fragment, constructing contiguous alignments of homologous reads and checking the alignments. It is also frequently necessary to reformat the alignments, loci or SNPs to be used in subsequent data analyses. There are many programs that can be used for these steps (Table 1.1).

Developing open source tools for researchers who do not have dedicated computational staff is an important endeavor for modern bioinformaticians. It is particularly important that these tools be user-friendly and easy to use and install. Multi-locus genetic data are the gold standard for phylogenetic and phylogeographic inference – facilitating the use of these data is a fundamental goal of my research. Chapters 2 and 3 describe computer programs I have written for this purpose.

Table 1.1. List of popular programs for quality control (Q), aligning (A), allele calling (AC), SNP calling (S) and visualization (V) of HTS data; whether the program is open source (O) and appropriate computer platforms also given.

Program	Q	A	AC	S	V	O	Computer Platforms	References
CLOTU	X	C				Y	Internet	Kumar et al. 2011
Galaxy	X	R		X		Y	Internet	Goecks et al. 2010
DNAstar SeqMan Ngen	X	R,D	X	X	X	N	Windows, MacOSX, Linux	http://www.dnastar.com
CLC Genomics Workbench	X	R,D		X	X	N	MacOSX, Linux, Windows	http://www.clcbio.com/
Geneious	X	R,D		X	X	N	WindowsVista, MacOSX	http://www.geneious.com/
GATK	X		X	X	X	Y	MacOSX, Linux	DePristo et al. 2011
RDP Pyrosequencing Pipeline	X					Y	Internet	Cole et al. 2009
Mothur	X					Y	Windows, MacOSX, Linux	Schloss et al. 2009
STACKS		C	X	X		Y	Unix	Catchen et al. 2011
CAP3		C				Y	Windows, MacOSX, Linux, Solaris, Internet	Huang and Madan 1999
PRGmatic		C,D	X	X		Y	MacOSX	Hird et al. 2011
ABySS		D				Y	Any	Simpson et al. 2009
SAMtools		R		X	X	Y	Unix	Li et al. 2009
BWA		R				Y	Any (C++ source)	Li & Durbin 2009
Bowtie		R				Y	Windows, MacOSX, Linux	Langmead et al. 2009
Exonerate		R				Y	Unix	Slater and Birney 2005
Novocraft		R				Y	MacOSX, Linux	Hercus 2009.
Stampy		R				Y	MacOSX, Linux	Lunter & Goodson 2011
SOAP		R,D		X		Y	Any (C++ source)	Li et al. 2008
MIRA		R,D		X		Y	MacOSX, Linux	Chevreux et al. 1999
Velvet		R*,D				Y	MacOSX, Linux, cygwin	Zerbino and Birney 2008
Bambino				X	X	Y	Windows, MacOSX, Linux	Edmonson et al. 2011
VarScan				X		Y	Any (JAVA source)	Koboldt et al. 2009
Casava				X		N	Linux	Illumina propietary
Tablet					X	Y	Windows, MacOSX, Linux, Solaris	Milne et al. 2009

R = reference; D = de novo; C = cluster generation

*velvet can use reference reads but it treats them as "just another" read, not a reference

1.3. HOST-ASSOCIATED MICROBIOTA

One of the major advances facilitated by HTS is in microbial ecology. For many years, sequencing of microbes required culturing a microbial species. However, only 1% of microbial species are believed to be culturable with contemporary techniques (Hugenholtz et al., 1998), although this number may be substantially higher (10%-50%) for gut microbes (Zoetendal et al., 2004). Advances in molecular methods eliminated this requirement and HTS allowed the sequencing of unprecedented microbial diversity in environmental samples from soil, water, the atmosphere and many locations on living hosts. The functions that microbiota provide for their hosts are numerous, ranging from simple competitive exclusion of pathogens (Sekirov et al., 2010) to bioluminescent camouflage (Nyholm & McFall-Ngai, 2004). The discovery of virtually unprecedented microbial diversity living on and in humans spurred research into gut microbiota – where it is estimated that over 2200 species reside (Zhang et al., 2009) in greater than 10^{11} microbes per gram of intestinal material (Whitman et al., 1998). The various localities of our bodies (i.e., mouth, colon, nostril, dominant hand, non-dominant hand) house statistically distinguishable assemblages of microbes (Costello et al., 2009); the gut is the most densely populated and diverse of the host-associated microbiota (Sears, 2005).

The functions provided by the gut microbiota in particular are as varied as mate selection (Sharon et al., 2010) and brain development (Heijtz et al., 2011). The gut microbiota is structured by host phylogenetics (Ley et al., 2008a), host ecology - specifically dietary specialization (Muegge et al., 2011) - but is also influenced by the environment (Benson et al., 2010), microbial interactions (Denou et al., 2009) and host genetics (Bevins & Salzman, 2011). Because of the close relationship to host genetics and environment, it is worth exploring the possibility that the vast genetic resource that is the gut microbiota may contain information on the shared evolutionary history of host and microbes. Quantifying the explicit roles of the above contributing factors is a major question in microbial ecology. Chapters 4 and 5 explore the various intrinsic and extrinsic factors associated with avian gut microbiota.

1.4. THE AVIAN DIGESTIVE TRACT (STUDY SYSTEM)

The gut, as an organ, predates many anatomical features of vertebrates – including bone and muscle (Stainier, 2005). Its primary and ancient function is to process food and extract energy for the organism. In birds, food enters through the mouth and passes immediately into the esophagus (Fig. 1.2). Birds frequently have a food-storing crop, followed by a stomach that potentially has two portions, the anterior proventriculus and the posterior gizzard. The proventriculus excretes hydrochloric acid and pepsinogen, to begin chemical digestion of the food. Birds lack teeth (presumably an adaptation for efficient flight) and the gizzard mechanically grinds the food using ingested stones before passage to the small intestine. At the junction of the small intestine with the large intestine is the cecum, of which there are usually two (Vispo & Karasov, 1996). The ceca vary in size and function, depending on the diet of the bird and function as the primary avian fermentation chambers (Józefiak et al., 2004). Posterior to the ceca is the large intestine, where any remaining water is absorbed, followed by the cloaca.

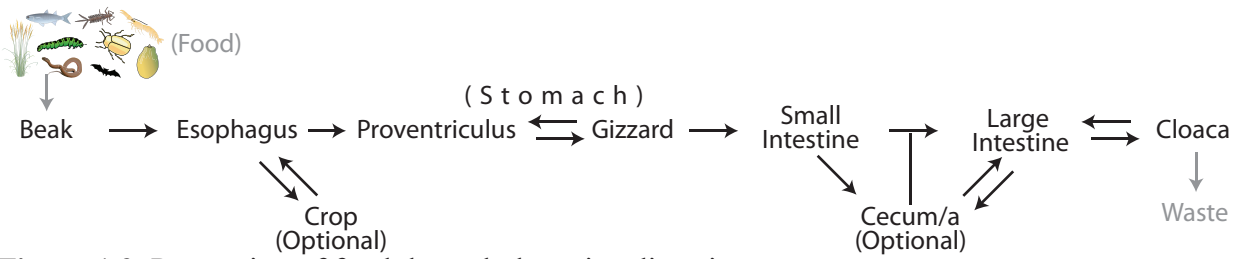


Figure 1.2. Procession of food through the avian digestive system.

The size and functions of these organs vary greatly across Aves (Fig. 1.3). Meat-eating birds tend to have reduced ceca and relatively simple digestive tracts (Fig. 1.3A). Birds that consume mostly seeds (or other hard materials) have an enlarged gizzard for pulverizing the food mechanically. The folivorous hoatzin has an enlarged crop (Fig. 1.3C) with complex microbial communities that degrade the leaves that comprise its diet. Other mostly plant consuming birds have elongated small or large intestines, to increase the amount of time and surface area the digesta is exposed to. The ceca has been the focus of many microbial studies, as its complex microbial communities are integral for digesting plant material. However, since not all birds have ceca and the ones that do can vary quite a bit in structure and function, we focused on the posterior portion of the large intestine for the gut microbiota chapters. The large intestine contains vast microbial communities that perform a relatively constrained function. The progression of foodstuff through the avian digestive system is not necessarily unidirectional (Fig. 1.2).

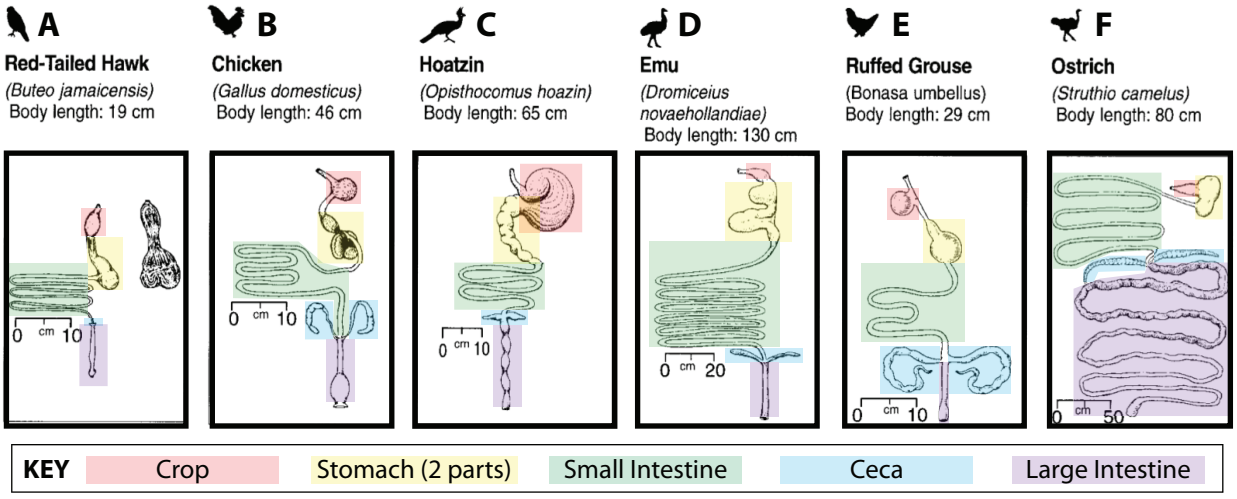


Figure 1.3. Comparative digestive anatomies of a meat-eating bird (A), an omnivorous bird (B), a folivorous bird (C) and three other plant-eating birds (D,E,F). The crop (red), stomach (yellow), small intestine (green), ceca (blue) and large intestine (purple) are color coded to show homology. Modified from Stevens and Hume (1995); used by permission from Cambridge University Press.

1.5. OVERVIEW OF CHAPTERS

This dissertation integrates all of my research interests into four research chapters (Fig. 1.1). First, I wrote a computer program that takes HTS data directly from the sequencer and reformats homologous reads into diploid loci (Chapter 2). This program, PRGMATIC, makes it possible for

a researcher to transform hundreds of thousands of sequencing reads to alleles that are suitable for evolutionary analyses, including species tree estimation (as in McCormack et al. (2012)), population genetics (as in Zellmer et al. (2012)) and hybrid zone analysis (as in Maley and Brumfield (2013)). The PRGMATIC README (Appendix A) walks through the steps and the PRGMATIC Guide to Common Errors (Appendix B) explains some potential issues that arise with restriction digest data and anonymous genomic fragments. These two documents are intended to be very user friendly and helpful for researchers with no interest in bioinformatics but who want their data to be appropriate for the questions they are asking.

LOCINGS (Chapter 3) allows the user to display summary information about the loci called from PRGmatic (or other programs that call diploid genotypes from HTS alignments). Certain information about HTS loci is required to deduce the quality of the raw data yet this information is frequently difficult to export from the most commonly used HTS processing programs. LOCINGS is lightweight, easy to install, and quickly displays the most important parameters for multi-locus, anonymous genetic loci: coverage and number of SNPs. The program also displays summarized information about both the loci (i.e., number of individuals called) and the individuals (i.e., number of loci called). These data can be exported as a table and the underlying sequence data can be output by locus or reformatted for input into several common evolutionary analysis programs. Again, the README (Appendix C) is specifically intended to be user-friendly and of use to researchers with no bioinformatics training.

The latter half of my dissertation focuses on evolutionary signal found in the bacteria living in the large intestine of birds. First, I use a brood parasite, the brown-headed cowbird, *Molothrus ater*, to investigate the role of nature and nurture in structuring the gut microbiota (Chapter 4). Since brood parasites deposit their eggs in the nests of other species instead of raising their own young, they provide a perfect natural system to elucidate the effects of genetics vs. rearing environment. Second, I specifically look for phylogeographic signal within the gut microbiota of 59 Neotropical bird species (Chapter 5). The samples span the central mountain ranges of Costa Rica and include seven individuals from Peru. In this chapter, I test whether the gut microbiotas have significant associations with an array of categorical and continuous variables, bearing on host taxonomy, host ecology, physical space and individual properties of the bird.

Finally, I synthesize the results of the above research chapters (Chapter 6) and discuss their implications and future directions.

Chapter 2.
PRGMATIC: An Efficient Pipeline For Collating Genome-Enriched Second-Generation Sequencing Data Using A "Provisional-Reference Genome"*

2.1. INTRODUCTION

Many research laboratories are interested in harnessing the sequencing capacity of second-generation sequencing (SGS) platforms, but are hindered by the lack of easily implemented bioinformatics tools for post-sequencing processing. Advances with genome enrichment techniques (or genomic reduction techniques, i.e., CRoPS (van Orsouw et al., 2007), modified AFLP protocols (Gompert et al., 2010, Zellmer et al., 2012, McCormack et al., 2012), RAD tags (Baird et al., 2008), molecular inversion probes (Absalan & Ronaghi, 2007), on-array and in-solution hybrid capture methods (reviewed in (Mamanova et al., 2009)) allow researchers to generate datasets that contain loci sampled from across the genome while maximizing overlap of these loci across individuals. The benefits from multi-locus, multi-individual sequence data are obvious – many questions in evolutionary biology (and other biological fields) require such data. Datasets such as this are attractive to researchers doing population or species level studies since many loci from multiple individuals provide improved inference compared to fewer loci and fewer individuals (Brumfield et al., 2008). Enriched genomes are also useful for identifying genomic areas of interest (i.e., under selection, highly variable, conserved enough for interspecific primers, etc.). However, SGS data from genome enrichment techniques pose some specific organizational and analytical problems and have thus far required each researcher to create, de novo, a set of bioinformatics tools to process them. Software that allows any researcher to collate these data into a common format (e.g., FASTA) will facilitate the evaluation of their quality and suitability for further analyses.

Here we outline a pipeline (PRGMATIC, for Provisional-Reference Genome automatic pipeline) for the analysis of SGS reads from the enriched genomes of individuals (in this case restriction enzyme-digested, size-selected genomic DNA sequenced on a Roche 454 system; a modified AFLP protocol similar to Gompert et al. (2010) except with individuals as the tagged units, instead of pooled population samples). PRGMATIC was designed with these major goals in mind: (1) Familiar output format (FASTA). (2) Friendly to use. (3) Free. (4) Relatively fast. PRGMATIC source code and documentation are available at https://github.com/shird/PRGmatic or https://sites.google.com/site/sarahhird/project-code/prgmatic.

A full plate of 454 pyrosequencing using a typical genome reduction method may consist of hundreds of thousands of reads from thousands of loci across tens of individuals (e.g., Gompert et al. 2010). To evaluate data such as these the pipeline constructs a "provisional-reference genome" (PRG) from the loci targeted through genome enrichment. First, reads within individuals are clustered at a high level of stringency (99% identity) into "alleles" (Figure 2.1). Second, the "alleles" are clustered across all individuals at a lower percent identity into "loci".

* This chapter previously appeared as: Hird S, Brumfield RT, Carstens BC. 2011. PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a "provisional-reference genome". *Molecular Ecology Resources*, 11(4): 743-748.

Third, a consensus sequence is created from each "locus" and all consensus sequences are concatenated to form the PRG. All original reads are then aligned to the PRG and individual genotypes can be called based on the cumulative reads that "stick" to each locus. The PRG is particularly useful because a plethora of analytic tools have been written for researchers who have a reference genome, yet relatively few researchers have access to a fully annotated "real" reference genome. This pipeline creates what can be used as a real reference genome but only contains information from reads contained within a sample. The nine discrete steps of PRGMATIC are outlined below.

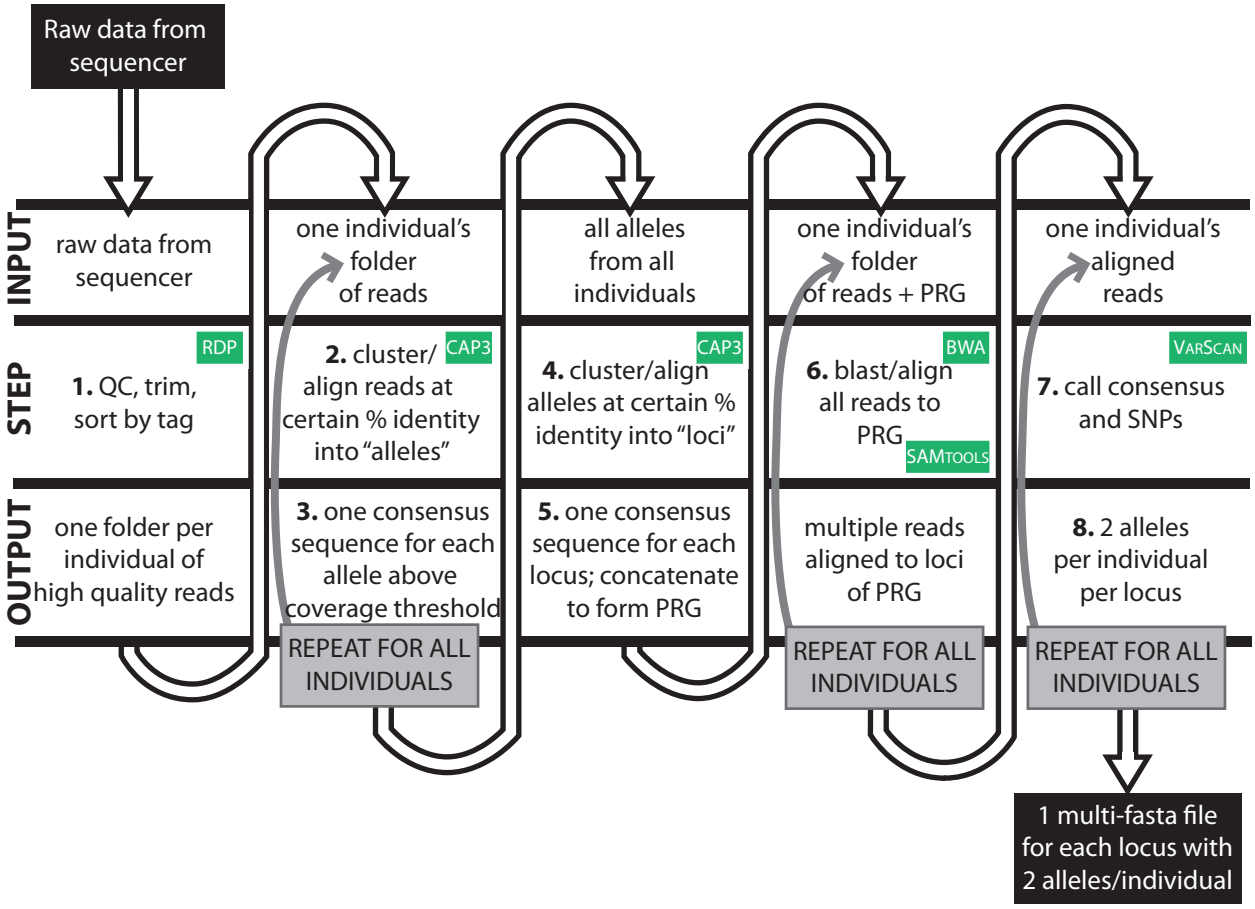


Figure 2.1. Schematic flowchart of PRGMATIC. Green boxes indicate dependent software used to accomplish each step (descriptions in Table 2.1). Bolded numbers refer to the numbered steps in the chapter. PRG, provisional-reference genome. SNP, single nucleotide polymorphism.

2.2. PRGMATIC

2.2.1. Step 1: Preprocess data

PRGMATIC requires that individuals are the sequence-tagged units (i.e., one sequence tag for each individual). The data that come directly from the sequencer needs to be separated by tags and quality controlled to remove low quality sequences. One option for preprocessing the data is the Ribosomal Database Project (Cole et al., 2009) website which has a "Pyrosequencing Pipeline" that will process raw data, apply quality filters and return tag-separated files.

2.2.2. Step 2: Cluster reads within each individual

PRGMATIC begins by clustering and aligning reads within each individual at a high percent identity using the program CAP3 (Huang & Madan, 1999). This step collects all reads that are almost identical (i.e., separated by very few SNPs/errors) into a single contig to form a putative allele. The default percent identity for collating alleles is 99%, but the user can modify this value.

2.2.3. Step 3: Call alleles

Contigs within an individual that are above a given coverage (default: 5X) are designated high-confidence alleles. A consensus sequence is called for each high-confidence allele. Higher minimum coverage results in fewer alleles and fewer loci, although the user may have higher confidence in their alleles. Lower minimum coverage can result in many loci that correspond to very small clusters within a single individual (and are thus unusable for downstream analysis).

2.2.4. Step 4: Cluster alleles across individuals into loci

A second cluster/alignment step (again using CAP3), at a lower "locus-level" percent identity (default: 90%) collates all the high-confidence alleles into contigs treated as putative loci. This collects all variants of a locus together and ensures that only one sequence per locus is contained in the PRG.

2.2.5. Step 5: Construct provisional-reference genome

PRGMATIC uses the loci to write the PRG, where each locus is annotated as if it were a chromosome in a full reference genome. The construction of the PRG allows the utilization of software designed for projects with a reference genome; these tools allow alignments to occur very quickly by eliminating pairwise comparisons across all reads and instead simply checking for similarity to the reference.

2.2.6. Step 6: Align all reads to the provisional-reference genome

The program BWA (Li & Durbin 2009) is called to format the PRG and align all the reads to it. Since there is some variation across individuals in the number of reads, as well as the quality of reads, this enables loci that are sequenced above some threshold in certain individuals to be detected and genotyped in potentially all individuals. Each locus can then be interpreted on an individual basis to detect SNPs and errors and compute a consensus sequence.

2.2.7. Step 7: Call SNPs within individuals

SAMTOOLS (Li et al. 2009) is used to sort alignments generated by BWA and outputs a summary of each individual's reads in "pileup" format. A custom Perl script converts the pileup format to a table of counts for all the reads at each position in the PRG. This table is then used to infer consensus sequence and SNPs. The program VARSCAN (Koboldt et al., 2009) is used to find insertions and deletions between the sequence reads and the PRG. These are incorporated in the next step. The user is prompted for the values they would like to use for minimum consensus

coverage, SNP coverage and SNP percent composition (what percentage of total reads the SNP comprises).

2.2.8. Step 8: Write alleles

A custom Perl script evaluates whether an individual's set of reads at a locus meet the minimum coverage cutoff value and writes each acceptable locus as a FASTA file with two alleles per individual. To write an allele, a Perl script evaluates the bases at each position. The base with the highest percent composition is the consensus sequence and written as the first allele. If there is a second base at a position that exceeds the SNP cutoff value and composition percent, the SNP is incorporated on the second allele, which is otherwise identical to the first. The cutoff values ensure that the SNP is both absolutely and relatively supported. (If 4X coverage and 20% read composition is required, a SNP with 5X coverage would be accepted if there were 10X total coverage at that base (50%), but not if there were 100X coverage (5%)). The unaligned loci files are output to a separate folder ("calledAlleles") in the PRGMATIC folder. The program MUSCLE (Edgar, 2004) may be used to align the called alleles; MUSCLE is a fast multiple sequence aligner that is not required for the functioning of PRGMATIC but provides aligned FASTA files, an input format necessary for many downstream sequence analyses.

This method of allele calling is heuristic, as it is entirely based on the reads in the sample and their relative composition to the sample as a whole; a more sophisticated method for allele calling, such as using maximum likelihood (Hohenlohe et al., 2010) that uses statistical models for allele inference, will be incorporated into future versions of PRGMATIC. Currently, statistical tests may be conducted on the data used to generate our allele calls, as they are output as separate files. However, with our current methodology, it is recommended that some loci are confirmed by the user by inspecting the raw data that generated the alleles (discussed further below).

2.2.9. Step 9: Compute summary statistics

The final step of the pipeline computes several summary files. The first is a table of how many and which individuals have been called for each locus. The second is a table of the number of individuals, number of heterozygotes, number of alleles, and the observed and expected heterozygosity for each locus. A third file contains information about multiple hits: if an individual has more than two bases for a single position, the individual, position and locus are recorded. Finally, if the COMPUTE portion of the ANALYSIS package (Thornton, 2003) is available on the local machine, a suite of summary statistics is computed for each locus (Watterson's θ, Tajima's D, etc.).

2.3. FOUR DEPENDENT PROGRAMS

CAP3 (Huang & Madan, 1999) is a fast sequence assembly program that incorporates base quality and automatically clips the low-quality ends of reads (Table 2.1). The clipping is controlled by several parameters that can be set by the user. CAP3 was chosen for the pipeline because it is quick, Unix-based, easy to use and free. It also automatically computes the consensus sequence for any contig it builds and outputs files in .ace format, which allow the contigs to be viewed in visualization programs like TABLET (Milne et al., 2009) or CONSED (Gordon et al., 1998).

By constructing a reference genome from the data, we are able to utilize several programs that have been designed for genome assembly. After clustering and assembling the reads into loci, and combining the loci into a PRG, BWA (for Burrows-Wheeler Aligner (Li & Durbin, 2009)) aligns all the reads within an individual to the reference. The advantages of BWA include speed and accuracy with reads greater than 200bp. After indexing the reference genome, BWA sequentially finds the starting position for each read in the reference genome. It then generates alignments and outputs in SAM format.

Table 2.1. Synopsis of software used by PRGmatic.

Program	Use	Internal dependencies	Required?	Citation
BWA	Quickly align reads to PRG	None	Yes	Li & Durbin 2009
CAP3	Cluster and alignment of reads	None	Yes	Huang & Madan 1999
COMPUTE	Calculate summary statistics of loci	libsequence	No	Thornton 2003
MUSCLE	Multiple sequence alignment	None	No	Edgar 2004
SAMTOOLS	Format/index reads aligned to PRG; pileup format	None	Yes	Li et al. 2009
TABLET	Visualization of SGS data	None	No	Milne et al. 2009
VARSCAN	Call consensus and SNPs from pileup format	None	Yes	Koboldt et al. 2009

SAM-formatted alignments are read by SAMTOOLS (Li et al., 2009) and converted to a binary version of the SAM format for quicker analysis; SAMTOOLS sorts the data according to their position in the reference genome and creates an index for both the reference sequence and the aligned reads to optimize speed. The reads are then compiled into pileup format, in which each line represents a reference base containing information on the number of reads, number of each base, read qualities, etc.

The pileup file is used to generate the consensus sequence and SNPs and is used as input for VARSCAN (Koboldt et al., 2009), which identifies insertions and deletions (indels).

2.4. TO USE PRGMATIC

PRGMATIC was developed for MacOSX. The user needs an "inputFASTA" folder of their data separated by tag (i.e., one FASTA file for each individual). All data must first be quality controlled to the level desired by the user and in multi-FASTA format. Once all the dependent programs have been unpacked and the executables of the dependent programs are placed in the PRGMATIC folder (accomplished using the included Setup script), the user enters a single command and is prompted for 5 parameters. If desired, the user can set additional parameters by opening the PRGMATIC script (written in Perl) and manually adding the appropriate flags to the script. This should not be necessary for the majority of cases, as we have prompted the user for the most influential parameters.

Upon completion of the script, the user may view all the contigs and the PRG with the aligned reads from each individual. This is useful for examining the quality of the data underlying the output, understanding how the programs work and discarding paralogous loci.

PRGMATIC was designed with speed intended as one of its primary strengths. We have tested the pipeline on datasets generated by several researchers. The fastest runtime was <20 minutes in which 780 loci were generated from a 20-individual dataset containing a total of 157,000 high quality reads. The longest runtime was 14 hours in which 545 loci were generated from an 80-individual dataset containing a total of 404,000 high-quality reads. (All preliminary data generated on a 2.66GHz Intel Xeon processor with 16 GB memory and used the default settings.)

2.5. PROOF OF CONCEPT

Two simulations and confirmation of 4 loci with Sanger sequencing provide proof of concept for PRGMATIC. The first included simulating 454 data using 100 empirical loci from a beta tester as a template with the program FLOWSIM (Balzer et al., 2010). We simulated 10,000 reads then quality controlled them by removing reads <100 bp and any reads containing an "N". This resulted in 6825 high quality reads that we then used as input for PRGMATIC. After <3 minutes of run time, 589 alleles (average coverage 10×) were identified; the 236 alleles with ≥ 5× coverage were subsequently clustered into 101 loci, forming the PRG. All the original reads were then aligned to the PRG (average coverage 60×; range of coverage 7 – 83 reads). We assembled the 101 PRGMATIC loci with the 100 empirical loci using GENEIOUS (Drummond et al., 2010). For 99 loci, each empirical locus aligned to one PRGMATIC locus without a single SNP to differentiate the two. There was one instance of two loci being called by PRGMATIC from a single empirical locus – the first PRGMATIC locus was identical to the empirical locus and the second was shorter than the empirical locus by 35 bp and contained two SNPs (one adjacent to a 2 bp homopolymer and the second within a 4 bp homopolymer). In other words, PRGmatic recovered 100% of the loci (and a 99% 1:1 correspondence between empirical loci and estimated loci) but also called a single "incorrect" locus which differed from the "correct" locus by 37 bp. This error may be due to the fact that the simulated data was not identical to genome enriched SGS data in that we did not ensure the correct forward and reverse primers were on each read.

The second simulation contained 5 individuals with 5 loci each: one monomorphic locus, one polymorphic at a single site locus where each individual is a homozygote, one locus with heterozygotes, one locus with a four base pair indel and a fifth "locus" with three unique alleles within 90% similarity of each other to simulate a paralogous locus (see Table 2.2). Data mimicked genome enriched data by containing primer sequences and individual sequence tags like each empirical read would have; then 10,000 reads per individual were simulated with FLOWSIM and edited by hand to remove all sequences in the reverse direction (which are not found in empirical datasets), data were sorted by tag and primer sequences were removed. Data were also quality controlled for sequences that were too short (<100 bp) or contained Ns. This resulted in an average of 2177 high quality reads per individual (range 2117 – 2245). The pipeline was run with the default settings – which took <20 minutes to complete. Six loci were called from 263 alleles – the simulated locus with three alleles was split into two loci whereas all the other loci were correctly identified. Additionally, all genotypes were correctly called with one exception: the four base pair indel, where heterozygotes for the indel were called with one allele containing three of the bases and the second allele containing the remaining base (see Table 2.2).

Finally, four loci have been empirically verified using Sanger sequencing (Maley & Brumfield, 2013). Primers were designed based on loci called by PRGMATIC and Sanger sequenced on an

ABI 3100; all were found to be single copy and the variation (1-4 SNPs/locus) identified by the pipeline were verified.

Table 2.2. Five individual (ind), five locus simulation conditions (ACTUAL) and results (RESULTS). Up to three alleles (X.1, X.2, X.3) are shown for each locus, the results contain the six loci called by PRGMATIC. The total length of each locus is given in parentheses. An asterisk denotes where PRGMATIC called the exact alleles as the simulated alleles. A dash denotes a gap.

		Monomorphic (302)		Polymorphic (288)		Heterozygous (331)		Indels (334/338)		Paralogous (323)			
		1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2	5.1	5.2	5.3	
ACTUAL	ind1	A	A	AGT	AGT	CA	CA	CACA	----	AACGC	AACGT	CTGTC	
	ind2	A	A	GGC	GGC	AT	CA	CACA	----	AACGC	AACGT	CTGTC	
	ind3	A	A	GGT	GGT	AT	CT	CACA	----	AACGC	AACGT	CTGTC	
	ind4	A	A	ATT	ATT	AA	AT	----	----	AACGC	AACGT	CTGTC	
	ind5	A	A	ATT	ATT	AA	AT	CACA	CACA	AACGC	AACGT	CTGTC	
		1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2	5.1	5.2	6.1	6.2
RESULTS	ind1	*	*	*	*	*	*	CAC	A	*	*	*	*
	ind2	*	*	*	*	*	*	CAC	A	*	*	*	*
	ind3	*	*	*	*	*	*	CAC	A	*	*	*	*
	ind4	*	*	*	*	*	*	*	*	*	*	*	*
	ind5	*	*	*	*	*	*	*	*	*	*	*	*

* Estimated genotype was identical to simulated genotype in column above

2.6. RECOMMENDATIONS

Perhaps the biggest consideration with the use of genomic reduction data is the identification of multi-copy genes. If a gene duplication event occurred recently, such that the two paralogs share a percent identity above the threshold for clustering alleles into loci, all the reads generated from the two different places in the genome will align to the same locus in the PRG. Grouping multiple loci into a single locus is problematic for SNP calling and may distort downstream analyses. It should, however, skew certain summary statistics in predictable ways. If paralogous loci have acquired high frequency or fixed differences, this should dramatically skew the apparent heterozygosity within populations. For this reason, a table containing observed and expected heterozygosity, calculated on a haplotypic basis, is included. If paralogs are grouped as one locus, the observed heterozygosity should be high relative to expected heterozygosity. The user would want to visually inspect their data as well as the summary statistics for loci that look biologically suspect. Actually viewing the reads that were used to call each locus should increase user confidence that a locus is homologous across individuals and across reads within a single individual. A supplementary guide to detecting common errors, complete with screen shots of various errors and real data, is included in the PRGMATIC distribution, in order to facilitate understanding of PRGMATIC output.

Chapter 3.
LOCINGS: A Lightweight Alternative For Assessing Suitability Of Next-Generation Loci For Evolutionary Analysis*

3.1. INTRODUCTION

To apply the immense sequencing capabilities of next-generation sequencing (NGS) technologies to population-level questions (i.e., those that require multi-locus, multi-individual data), genome enrichment methods are frequently employed. These methods aim to sample the genome at a reproducible subset of markers that can be obtained from many individuals and reduced to genotype (i.e., a set of phased alleles). Examples of these methods include amplicon sequencing (Binladen et al., 2007), RAD-tags (Baird et al., 2008), complexity reduction of multilocus sequences or CRoPS (van Orsouw et al., 2007) and sequence capture (Okou et al., 2007); for a review of NGS methods suitable for multi-locus studies, see (McCormack et al., 2013) . Genome enrichment methods often utilize a known or constructed reference for easing alignment of sequencing reads. Genotypes can then be called from the alignments, using a variety of bioinformatics methods (e.g., (Hird et al., 2011b), Catchen et al. (2011). This results in next-generation alignments to a reference and a set of loci for the individuals in the study; the loci can then be used in standard phylogeographic, phylogenetic or population genetic studies or other multi-locus analyses (e.g., McCormack et al. (2012), Zellmer et al. (2012)). Prior to analysis, however, researchers must determine which loci are suitable for the questions being asked by assessing key parameters such as coverage and number of polymorphic sites or whether all populations are represented.

Current NGS file types are efficient at manipulating and storing alignment data but the parameters of interest are difficult to extract and can require custom bioinformatics scripts. Additionally, these file types are not useable in downstream analyses. Although large-scale, comprehensive programs like the Genome Analysis Toolkit (GATK) (McKenna et al., 2010) can calculate coverage, if the parameters of interest are limited and include coverage per locus and coverage per individual, these programs are more heavy-duty and time-intensive than a user may want to invest. LOCINGS is a lightweight, easy to use program that displays and outputs key parameters for researchers interested in multi-locus analysis of genotypes.

As more NGS papers are published, it should be standard to report summary statistics about coverage and polymorphism, in addition to the already standard number of total and high quality reads. Furthermore, as sequencing capacity continues to increase, the number of loci and number of individuals in a dataset will as well. Easily accessing, summarizing and reporting these parameters are important steps toward streamlining analysis and understanding large multi-locus datasets. LOCINGS does not analyze any of the user-supplied data – it simply reports and exports summarized information about the dataset contained in the input files that is difficult to extract manually.

* This chapter previously appeared as: Hird SM. 2012. lociNGS: a lightweight alternative for assessing suitability of next-generation loci for evolutionary analysis. *PLoS ONE*: 7(10): e46847.

3.2. METHODS

3.2.1. Overview

LOCINGS was designed for use with multi-locus, multi-individual datasets generated through NGS. It collates information about loci, alignments and demographic data so that users can view summarized information about the genetic data (Table 3.1; Fig. 3.1) on the same screen as taxonomic and field data (e.g., subspecies, sampling locality, gender, etc.). In this way, one may assess the suitability of the data for further analysis.

A

lociNGS

Individual	Population	numLoci	totalReads	usedReads	percentUsed
R01	POP1	165	7261	2756	37
R02	POP1	45	4307	1335	30
R03	POP1	56	4129	1462	35
R04	POP1	151	7447	2582	34
R11	POP2	263	10861	3828	35
R12	POP2	210	7845	3122	39
R13	POP2	437	14481	5807	40
R14	POP2	257	9525	3443	36

B

R03

Locus Name	Length	SNPs	Coverage_This_Ind	Number_Inds	Coverage_Total	Coverage_Used
maleyData_118	333	5	8.0	5	71.0	64.0
maleyData_124	290	0	6.0	4	38.1	32.1
maleyData_125	322	3	8.0	7	143.9	139.0
maleyData_134	367	9	8.0	6	71.1	61.1
maleyData_140	298	1	6.0	4	55.0	41.0
maleyData_141	341	0	10.0	4	42.0	32.0
maleyData_142	320	7	8.0	6	64.6	59.599999999999994
maleyData_143	294	4	8.0	5	60.0	54.0
maleyData_144	302	6	9.0	6	88.0	79.0
maleyData_145	354	1	6.0	4	45.0	35.0
maleyData_146	316	4	8.0	4	45.0	33.0
maleyData_151	338	1	6.0	4	42.0	28.0
maleyData_152	335	4	8.0	5	62.0	52.0
maleyData_154	364	1	9.0	3	41.0	27.0
maleyData_157	338	0	8.0	2	30.0	17.0
maleyData_159	302	4	9.0	3	46.0	30.0

Figure 3.1. Screen shots of LOCINGS. Data include 8 individuals (rails); summarized data for the whole dataset shown in the summary screen (A) and one example of an individual (R03) screen shows parameters associated with individuals (B). Details of the column headings are in Table 3.1.

The program has two types of display screens, both in table format. The "summary screen" contains demographic data, number of loci per individual (numLoci), total number of reads sequenced, number of reads used (along with the percentage of total). The numLoci data serve as buttons that open the corresponding "individual screen". This screen displays specific information about all the loci found in an individual, including length of the locus, number of polymorphic sites, number of individuals sequenced for that locus and coverage (for the individual, for all individuals, and for only the individuals with high enough coverage to be called). Each of the coverage categories serves as buttons that print the corresponding raw data in multi-FASTA format.

Table 3.1. LOCINGS parameters for the summary screen (Sum; Fig. 3.1a) and the individual screen (Ind; Fig. 3.1b).

Screen[a]	Parameter[b]	From[c]	*[d]	Definition
Sum	Individual	Demo		The individual's name
Sum	Population	Demo		The individual's population of origin
Sum	numLoci	Align	*	The number of loci called for each individual
Sum	totalReads	Align		Total number of reads sequenced in each individual
Sum	usedReads	Align		Total number of reads used for calling loci in this individual
Sum	percentUsed	lociNGS		UsedReads/TotalReads
Ind	LocusName	Loci		The name of the locus
Ind	Length	Loci		Number of bases in the locus
Ind	SNPs	Loci		Number of polymorphic sites
Ind	Number_Inds	Loci		Number of individuals called for this locus
Ind	Coverage_This_Ind	Align	**	Coverage for this locus in this individual
Ind	Coverage_Total	Align	**	Total coverage across individuals for this locus
Ind	Coverage_Used	Align	**	Total coverage for all individuals used in final locus

[a] Which screen the data are displayed on, the summary or the locus screen

[b] Column header displayed in program; see Figure 1

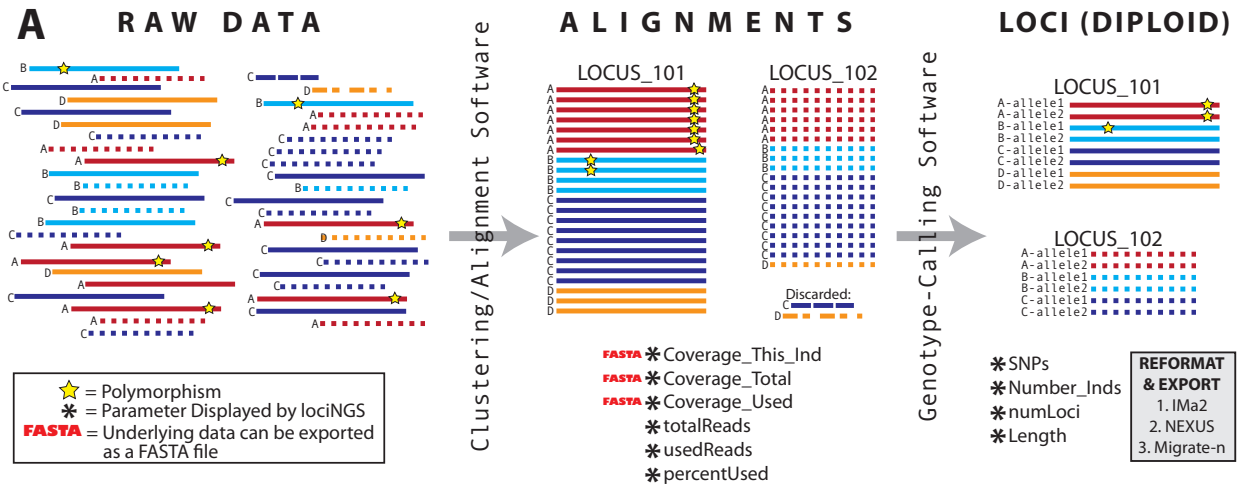
[c] Which input file the data are derived from, demographic data (demo), SAM/BAM alignments (Align), multi-FASTA locus files (loci) or calculated by lociNGS

[d] * indicates this column's data serve as a button to pull up locus screen; **indicates this column's data serves as a button to print the corresponding reads to a multi-FASTA file

3.2.2. Program Input

LOCINGS takes three categories of input: NGS alignment files, locus files (Fig. 3.2) and a demographic data file. When using genomic enrichment methods (or genome assembly methods), an alignment of the raw sequencing reads to a reference genome is often constructed using clustering or alignment programs, e.g., Geneious (Drummond et al., 2012), Galaxy (Goecks et al., 2010), Velvet (Zerbino & Birney, 2008), etc. One common format for these alignments is SAM (Sequence Alignment/Map (Li et al., 2009)) format or its binary version, BAM. These alignments contain a lot of information about the sequences and are LOCINGS's source for many of the coverage and sequence data parameters (see Table 3.1). For input to the

program, the alignment files need to be in sorted, indexed BAM format; the program SAMTOOLS (Li et al., 2009) can be used to convert SAM to BAM, sort and index the reads, if necessary.



B

	LOCUS 101	LOCUS 102
SNPs	2	0
Number_Inds	4	3
Length	200	150
Coverage_Total	23	19
Coverage_Used	23	18
Coverage_This_Ind A	7	6
Coverage_This_Ind B	4	3
Coverage_This_Ind C	9	9
Coverage_This_Ind D	3	1

C

	INDIVIDUAL A	B	C	D
totalReads	13	7	19	5
usedReads	13	7	18	4
percentUsed	100	100	94	80
numLoci	2	2	2	1

Figure 3.2. How the data are generated, where the parameters come from and example data. (A) Letters represent individuals and lines represent sequences; there are four individuals and two loci. Raw data from the sequencer is put through an alignment or clustering program to collect reads into alignments. From each alignment file, LOCINGS reports totalReads, usedReads, percent reads used (percentUsed), Coverage_This_Ind, Coverage_Total and Coverage_Used; LOCINGS will also export the data underlying the coverage parameters in FASTA format. Genotype-calling software will reduce sequence reads to loci (phased alleles). LOCINGS uses these loci to report SNPs, Number_Inds, numLoci and Length; the program can reformat the loci into IMa2, NEXUS or Migrate formats. For further explanation of the parameters, see Table 3.1. (B) The parameter values for the two loci (LOCUS_101 and LOCUS_102) in this example. (C) The parameter values for the four individuals (A,B,C,D) in this example.

Many traditional evolutionary analyses require individual loci that contain phased, homologous alleles for the individuals in the dataset. To get from alignments to loci, genotype-calling software is required, e.g. PRGMATIC (Hird et al., 2011b), STACKS (Catchen et al., 2011), GATK (DePristo et al., 2011, McKenna et al., 2010), etc. The loci are analogous to traditional Sanger sequencing loci and should be in multi-FASTA format. The locus files are the source for the SNP parameter as well as the locus names and length (see Table 3.1).

Finally, a demographic text file is required that, at a minimum, assigns each individual to a population; designating populations is frequently important in population level questions and is required because the output formats are capable of outputting a subset of populations or individuals. However, if this information is unknown or the user does not need the IMa2 or

migrate output options, population can be set to something meaningless and the program will function properly.

3.2.3. Program Output

LOCINGS outputs several different types of data. First, a table of all the information displayed to the user may be printed as a tab-delimited text file. This can then be edited with a spreadsheet or text-editing program to calculate averages, construct graphics, sort the data, etc.

Second, the raw sequences that were used to call a locus may be exported for an individual, for all individuals or just the individuals that were used in the final dataset; this information is contained in the alignment files but difficult to extract manually. These data are FASTA formatted.

Third, users may reformat a subset of populations or individuals into NEXUS (Maddison et al., 1997), IMa2 (Hey, 2010) or Migrate (Beerli & Felsenstein, 1999) formats. These three formats are highly specific and are used in population genetics programs that can analyze large, multi-locus datasets. In addition, these formats can be rather time consuming to produce by hand or require custom scripts to produce for more than a few loci. LOCINGS automates and combines the selection of loci and the construction of the appropriate input files. Under the export menu of the program, users select either populations or individuals they would like to include in the output of these formats; LOCINGS then searches all the loci that contain at least one individual from the populations selected or all individuals selected.

The location of all exported files is logged to the screen and each has a unique file name.

3.2.4. Test Data

There is a small test dataset provided with the LOCINGS distribution. This dataset includes four individuals at five loci. A copy of the exact parameter values displayed by LOCINGS with the test data is included as Table 3.2.

3.2.5. Program Implementation

LOCINGS is written in Python for a Unix-based system (e.g., MacOSX). It requires MongoDB as a separately installed program. LOCINGS uses the TKINTER class of Python for a user-friendly graphical user interface. A modified version of SEQLITE (available: http://www.mbari.org/staff/haddock/scripts/) calls polymorphic sites from the aligned locus files; this tool works by simply counting the variable sites in an aligned FASTA file. The BAM files are not considered in the number of SNPs. The User Manual is included as an Appendix (Appendix C).

3.3. AN EXAMPLE: USING LOCINGS IN PHYLOGEOGRAPHY

For many evolutionary analyses, a phased set of alleles is required as input; many NGS molecular and computational methods are now capable of producing such datasets. For example, McCormack et al. (McCormack et al., 2012) generated restriction-digested fragments sequenced on a Roche 454 platform for two species of rails (*Rallus longirostris* and *R. elegans*) to identify

fixed genetic differences in a bird hybrid zone; in this section I walk through a subset of their dataset that contains four individuals from each species (*R. longirostris* = R01, R02, R03, R04; *R. elegans* = R11, R12, R13, R14). The data was quality controlled and analyzed with PRGMATIC (Hird et al., 2011b), then loaded in to LOCINGS. The summary screen (Fig. 3.1A), which can be exported as a tab-delimited text file, informs the user of how efficient the method was, in terms of how many reads were aligned to the reference genome compared to total number of reads (Fig. 3.3). It also displays the total number of loci that each individual belongs to; these data functions as a button that opens the individual screen for the given individual (Fig. 3.1B).

Table 3.2. The exact parameters values output by lociNGS with the test data.

A. Summary Screen

Individual	Population	num Loci	total Reads	used Reads	percent Used	Longitude	Latitude	Location	Species
testA	POP1	5	3472	1272	36	-109.876	45.678	NoPlace, TX	Tamias bunkus
testB	POP2	4	1753	659	37	-109.876	45.678	NoPlace, TX	Tamias bunkus
testC	POP1	5	5138	1881	36	-109.876	45.678	NoPlace, TX	Tamias bunkus
testD	POP3	5	2139	593	27	-109.876	45.678	NoPlace, TX	Tamias tamias

B. Individual Screens

	Locus Name	Length	SNPs	Coverage_This_Ind	Number_Inds	Coverage_Total	Coverage_Used
testA	false_1	136	1	170.0	3	630.0	630.0
	false_2	145	0	31.0	4	81.0	74.0
	false_3	282	4	210.0	4	720.0	720.0
	false_4	191	1	250.0	4	761.0	761.0
	false_5	177	7	180.0	4	260.0	260.0
testB	false_1	136	1	170.0	3	630.0	630.0
	false_3	282	4	210.0	4	720.0	720.0
	false_4	191	1	250.0	4	761.0	761.0
	false_5	177	7	180.0	4	260.0	260.0
testC	false_1	136	1	170.0	3	630.0	630.0
	false_2	145	0	31.0	4	81.0	74.0
	false_3	282	4	210.0	4	720.0	720.0
	false_4	191	1	250.0	4	761.0	761.0
	false_5	177	7	180.0	4	260.0	260.0
testD	false_1	136	1	170.0	3	630.0	630.0
	false_2	145	0	31.0	4	81.0	74.0
	false_3	282	4	210.0	4	720.0	720.0
	false_4	191	1	250.0	4	761.0	761.0
	false_5	177	7	180.0	4	260.0	260.0

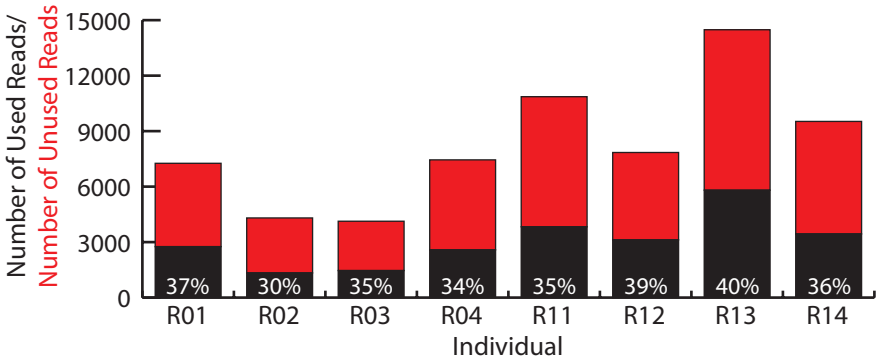


Figure 3.3. Number of reads per individual. Black portion of bars represents reads aligned to the reference; red portion accounts for unused reads. Percentage of reads used is shown in white text.

The individual screen contains detailed information about each of the loci with links to the raw data that make up each locus (Fig. 3.1B). Exporting this data as a tab-delimited text file allows the user to determine the distributions of polymorphic sites (Fig. 3.4A), number of individuals (Fig. 3.4B) and coverage per individual (Fig. 3.4C) across all loci. One can also assess how well each individual performed, by calculating average coverage. One may use this information to decide which individuals are worth resequencing with custom primers (to fill in their data matrix) or how to prune their dataset to the most complete or informative loci.

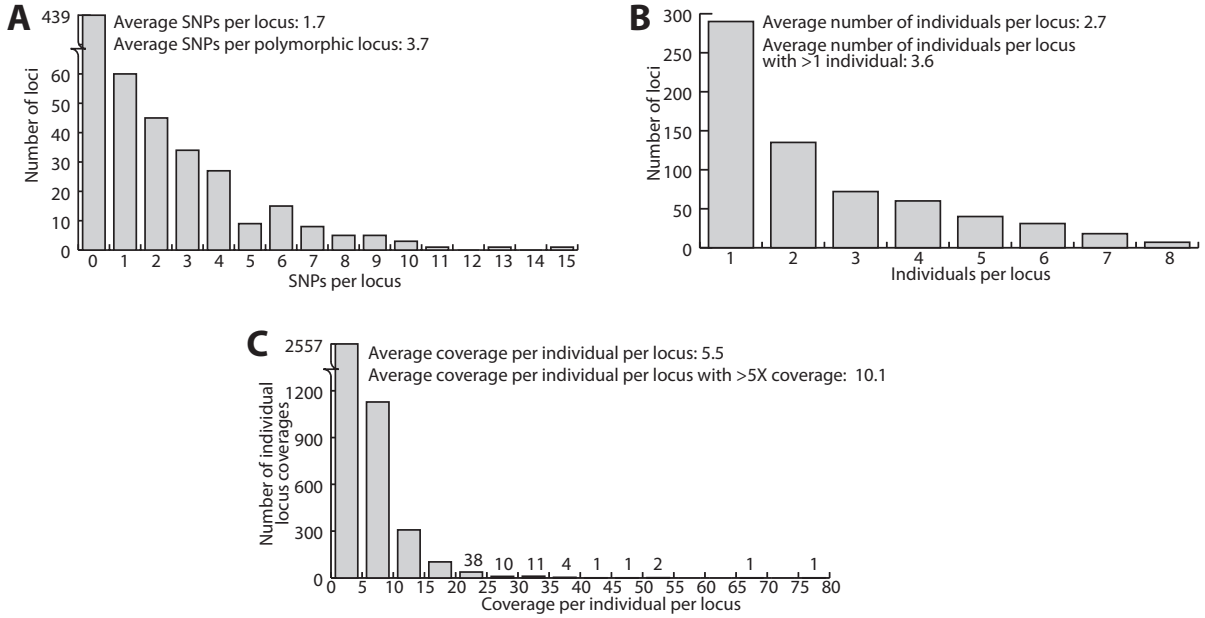


Figure 3.4. Summary histograms of important parameters in the rail dataset. Number of polymorphic sites (A), individuals present in each locus (B), individual coverage on a per locus basis (C). Note the scale of the dependent axis changes on (A) and (C).

If a particular locus has more polymorphic sites than one might expect by the processes of natural selection or drift, the user can output the sequence reads that compose the raw data to investigate underlying copy number. With the raw read data, an alignment and phylogenetic tree can be estimated from either a single locus for one individual or all the reads underlying a single locus from all individuals (Fig. 3.5), but analysis of the raw reads is up to the user. For these data, I used Muscle (Edgar, 2004) for alignment (using all defaults) and Geneious (Drummond et

al., 2012) to construct a neighbor-joining tree (using an HKY model of genetic distance and no outgroup). An analysis like this is very quick and although more sophisticated phylogenetic algorithms exist, for the purposes of assessing number of clades, these methods worked well. Once a tree has been constructed, if there are two (or fewer) major clades for each individual, it is likely that the sequences derive from a single diploid locus (Fig. 3.5A). However, if there are more clades than the ploidy of the organism allows, there may be multiple genomic sources of the data (Fig. 3.5B). One can also assess paralogy in the reads from all individuals at a locus: if all the reads from each individual belong to two or fewer clades, the locus is likely single copy (Fig. 3.5C). However, if one or more individuals belong to multiple clades, the underlying copy number may not be one (Fig. 3.5D).

Finally, LOCINGS will export the data in three formats for input to evolutionary analysis programs. Users select exportation of either individuals or populations. The program searches for all loci that contain at least one individual from each of the selected categories. In other words – if all individuals are selected, only the loci that contain all individuals will be reformatted and printed. If all populations are selected, only the loci that contain at least one individual from each population will be reformatted and printed.

Altogether, these simple functions provide the user with an overall sense of how their method and data perform at a basic level.

3.4. CONCLUSIONS

With the ever-increasing amount of data that is gathered with NGS, it is important to assess the suitability of the reads for further analysis. LOCINGS provides a simple and quick way to determine which loci and which individuals have enough coverage and polymorphism to use in evolutionary analysis. Furthermore, the program automatically converts suitable loci to several file formats that are common in evolutionary analysis and time consuming when done by hand. Small, easy to use programs designed for a specific task allow researchers to customize their workflow and minimize or eliminate the learning curve for complex programs.

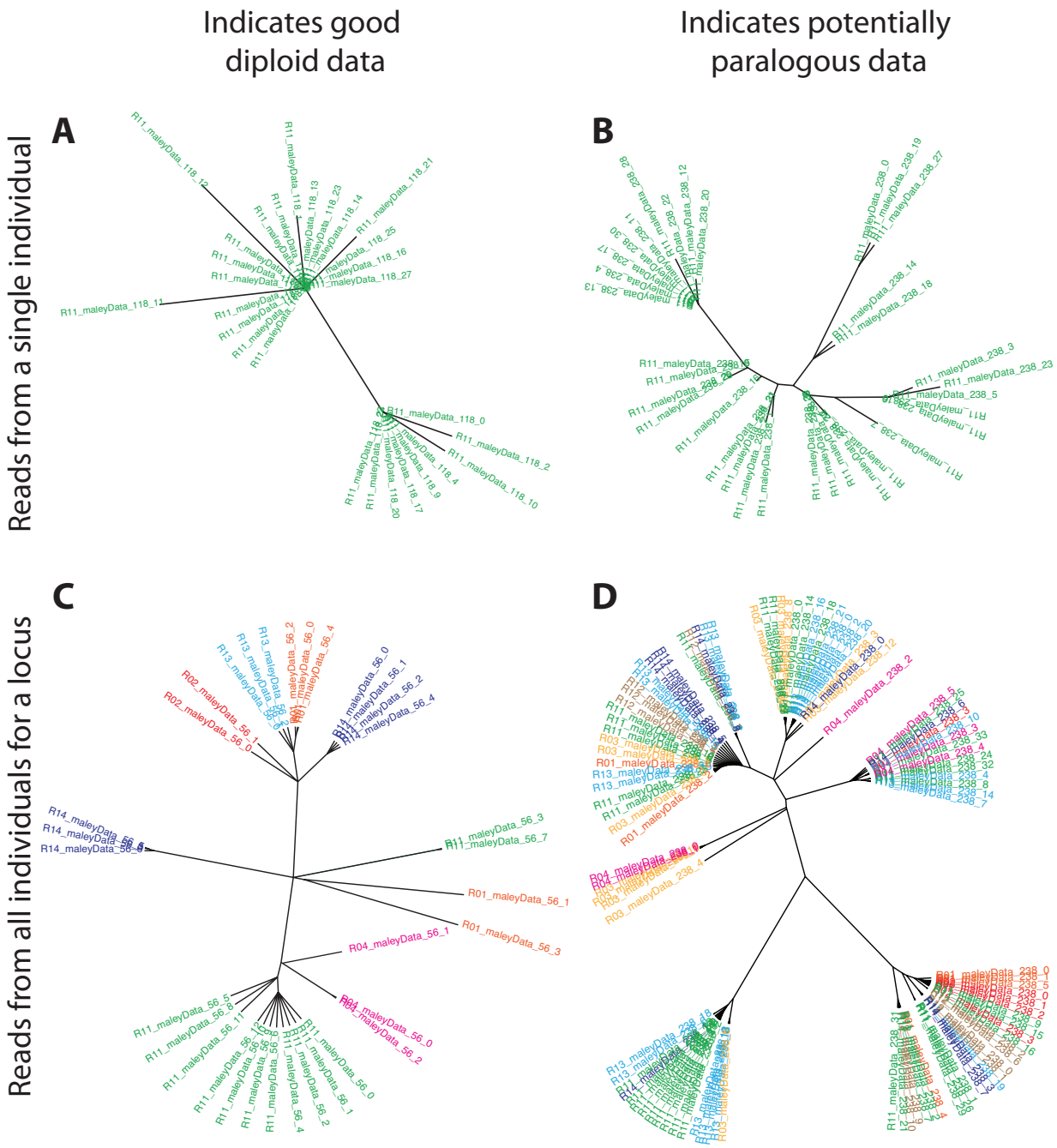


Figure 3.5. Neighbor-joining trees of aligned reads (reads output from the program) to help assess copy number. Shown are reads from one individual (A, B) and all the reads for a locus (C, D). Both (A) and (C) imply single copy loci; in (A) there are only two major clades and in (C) the reads for each individual, as shown by the different colors, belong to two clades at the most. Both (B) and (D) indicate potential multi-copy loci; in (B), there are greater than two clades and in (D) the reads for each individual, as shown by the different colors, are frequently distributed across greater than two clades.

Chapter 4.
Nature, Nurture And The Gut Microbiota Of The Brood-Parasitic Brown-Headed Cowbird (*Molothrus ater*)

4.1. INTRODUCTION

The gut microbiotas of vertebrates are complex communities of millions of organisms that influence a vast number of essential processes in the host, including mate choice (Sharon et al., 2010) and brain development (Heijtz et al., 2011). The number of genes in the gut microbiota can outnumber host genes by several orders of magnitude (Xu et al., 2003). The microbiota comprise a vast genetic resource that shares many of the evolutionary pressures of the host yet are also exposed to local selection pressures within the gut (Ley et al., 2006). Despite this biological importance, it is not well understood how the microbiota of a particular individual is determined (Benson et al., 2010).

A variety of factors are correlated with the structure of the gut microbiota, including the host genotype, host ecology and the local environment. Host genotype can directly influence microbiota membership (Petnicki-Ocwieja et al., 2009, Tims et al., 2011, Zoetendal et al., 2001). Genetic distance between hosts is correlated with microbiota similarity at several taxonomic scales, from within a single species (Yatsunenko et al., 2012) to across closely related species (Ochman et al., 2010), and up to higher taxonomic levels (Phillips et al., 2012, Ley et al., 2008a, Ley et al., 2008b, Colman et al., 2012). Host ecology also has an influence. In mammals, fecal microbiota of hosts of different dietary specializations (i.e., omnivore, carnivore, herbivore) cluster in multivariate space (Ley et al., 2008a) and the associations are stronger than phylogenetic effects (Muegge et al., 2011). Taxonomically and ecologically diverse insects show associations between the microbiota and both host taxonomy and diet (Colman et al., 2012), with the strength of the association varying by species. Local environment also contributes – for example, in mice, being raised together has a stronger effect on gut microbiota than being genetically related (Benson et al., 2010). One or more of these factors has led a diverse set of organisms to show population level structure in gut microbial communities, including hoatzins (Godoy-Vitorino et al., 2012), chimpanzees (Degnan et al., 2012), humans (Lee et al., 2011) and carnivorous plants (Koopman & Carstens, 2011). However, assigning this differentiation conclusively to genetics, ecology or environment is difficult, as physical space is usually positively correlated with genetic distance and accompanied by divergence in environment.

Brood parasites offer a unique natural system to investigate these processes that naturally separates vertical (genetic) and horizontal (environmental) transmission of microbes. Instead of building nests and raising their young, brood parasites lay eggs in the nests of brood host species, thus leaving brood host parents to spend reproductive resources on the parasitic young instead of their own babies (Fig. 4.1A). Brown-headed cowbirds (*Molothrus ater*, hereinafter cowbirds) are a generalist brood parasite; without any particular egg camouflage they parasitize at least 225 host species in North America (Lowther, 1993). They have evolved many adaptations for this lifestyle. An unusually good immune system prepares them for development in a variety of environments (Hahn & Smith, 2011), flexibility in egg laying behavior gives the cowbird control over when and where to lay eggs (Woolfenden et al., 2003), tolerance of host nestlings increases host parent feeding (Kilner et al., 2004), a relatively large gape width and quick growth rate

allow them to outcompete nest mates for resources (Ortega & Cruz, 1992) and thick eggshells protect the egg from puncture ejection by host parents (Spaw & Rohwer, 1987). Because the microbiota is critically important to the health of an organism, the communities of microorganisms may also show some degree of specialization to brood parasitism.

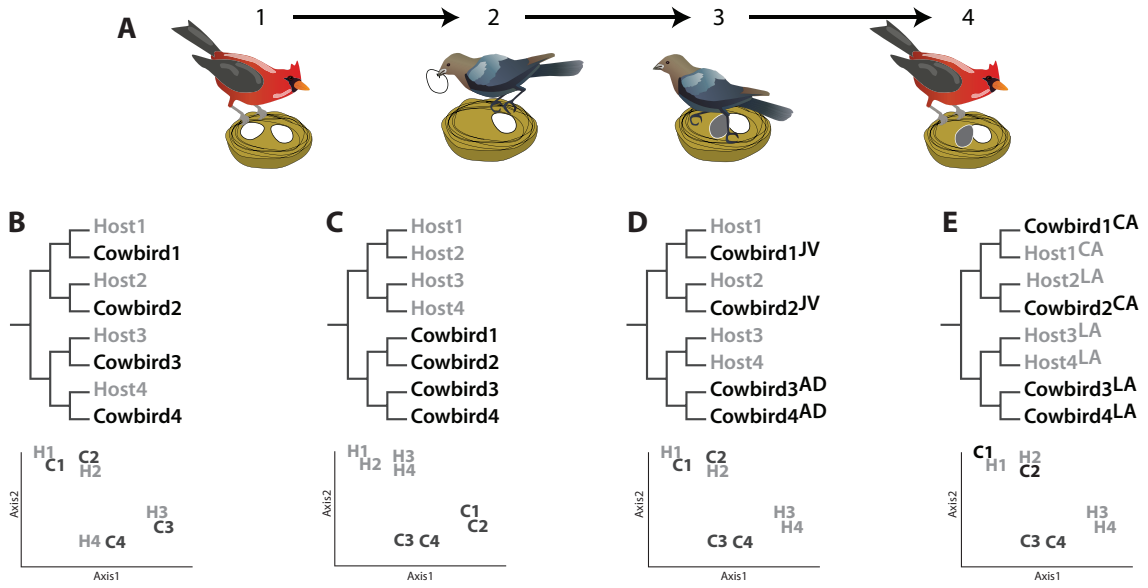


Figure 4.1. Cowbird reproductive strategy (top row), how the data could support the three major hypotheses through dendrograms (middle row) and principle coordinates analyses (bottom row) of four cowbird microbiota samples (C01-C04) and four host species (H01-H04). (A) 1. A host at its nest. 2. When host is gone, the cowbird ejects a host egg. 3. The cowbird lays one of its own eggs in the host nest. 4. The host returns to incubate and raise the cowbird alongside its own young. (B) Cowbird microbiota may most closely resemble their host microbiota, the "Nurture Hypothesis". (C) Cowbird microbiota may be most closely related to other cowbirds, despite their host species, the "Nature Hypothesis". (D) Cowbird microbiota may shift between being host-like when they are juveniles (JV) and cowbird-like when they are mature (AD), the "Convergence Hypothesis". (E) Local factors may determine the gut microbiota, causing birds from different localities (LA vs. CA) to be most similar.

Here we test what factors may shape the gut microbiota of cowbirds. We have four hypotheses for how the gut microbiota are structured (Fig. 4.1). First, the brood host may be the predominant influence on gut microbiota and this is supported if cowbird samples are more similar to potential brood host species than to other cowbird samples (the Nurture Hypothesis, Fig. 4.1B). Second, the process may be more static and cowbirds may have a cowbird-specific microbiota that is intrinsically (genetically) determined (the Nature Hypothesis, Fig. 4.1C). This is supported if the cowbird samples form a clade in which all cowbird samples are more similar to each other than to any of the potential brood host species. Third, adult and juvenile cowbirds may have different assemblages at different life stages (the Convergence Hypothesis, Fig. 4.1D). Specifically, the juvenile cowbirds may have a more general microbiota assemblage that is able to utilize a variety of diets but which converges to a stable cowbird microbiota as the cowbirds mature. This is supported if the microbiota of juvenile cowbirds are most similar to that of their brood host species, but the microbiota of adult cowbirds form their own clade. This may be the most likely hypothesis; both wild birds (Godoy-Vitorino et al., 2010) and domestic poultry

(Scupham, 2007, Scupham, 2009, Yin et al., 2009) undergo complex microbial maturation processes as they age. Mammals show a distinct transition of gut microbiota composition at the age of weaning (Inoue & Ushida, 2006, Simpson et al., 2000, Palmer et al., 2007). Finally, it may be that the local environment entirely accounts for similarity of gut microbiota – in this case, birds in closer geographic proximity will be most similar, despite their genetic background, ecology or evolutionary history (the Environment Hypothesis, Fig. 4.1E). Our null hypothesis is that there is no difference between cowbirds and brood host species. To put the variation in our samples in a broader context, we compare the birds to two large, phylogenetically distant groups – mammals and insects. A generalist gut microbiota may show higher levels of variation than other organisms with more specific diets.

4.2. MATERIALS AND METHODS

4.2.1. Sampling

We sampled birds available in the LSU Museum of Natural Science's Collection of Genetic Resources for this investigation, including 32 cowbirds (Icteridae: Molothrus ater) from two localities (Louisiana and California) and 16 individuals from nine known brood host species, also from LA and CA (Table 4.1, Appendix D): Northern Cardinal (Cardinalidae: *Cardinalis cardinalis*), House Finch (Fringillidae: *Haemorhous mexicanus*), Orchard Oriole (Icteridae: *Icterus spurius*), Indigo Bunting (Cardinalidae: *Passerina cyanea*), Blue-gray Gnatcatcher (Polioptilidae: *Polioptila caerulea*), Prothonotary Warbler (Parulidae: *Prothonotaria citrea*), Carolina Wren (Troglodytidae: *Thryothorus ludovicianus*), White-eyed Vireo (Vireonidae: *Vireo griseus*), Hooded Warbler (Parulidae: *Setophaga citrina*). Birds were frozen within two hours of collection. Throughout the manuscript, individuals are identified by the first letter of the genus followed by the first four letters of the species and an individual identifying number. One individual ("NorthernCardinal4") served as a replicate to assess PCR/sequencing bias; these samples are identified as NorthernCardinal4.1 (NC4.1) and NorthernCardinal4.2 (NC4.2) and bring the total number of samples to 49.

4.2.2. DNA Extraction, Amplification, Sequencing and Quality Control

The entire digestive tract was removed when the bird was thawed for museum specimen preparation. Total DNA was immediately extracted from the contents of the large intestine, halfway between the ceca and cloaca. Following Gloor et al. (2010) we used combinatoric primers and massive multiplexing of PCR amplicons for sequencing on an Illumina Hi-Seq. This method uses paired-end sequencing technology to generate pairs of sequences with 100% overlap across variable region 6 (V6) of the 16S component of rRNA; primer sequences align to positions 967-985 and 1078-1061 on Escherichia coli 16S rRNA (Gloor et al., 2010).

We used several measures of sequence quality control. First, both reads of a given pair had to match across 100% of the bases. The pairs also had exactly matching tag sequence and no errors in the priming sequence. We used the Bellerophon (Huber et al., 2004) function within the mothur program (Schloss et al., 2009) to identify and discard potentially chimeric sequences. Finally, we used mothur to discard sequences that did not blast to the domain Bacteria. The reads passing these filters were included in the final dataset.

Table 4.1. Number of individuals from each locality, adult (AD) / juvenile (JV) status and rate of brown-headed cowbird parasitism (Ortega 1998) for the species in this study.

Common Name (Abbreviation)	Scientific Name	California		Louisiana			Parasitism
		JV	AD	JV	AD	Totals	Rate (%)
Northern Cardinal (NC)	*Cardinalis cardinalis*	0	0	2	2	4	2.7 – 100
House Finch (HF)	*Haemorhous mexicnus*	1	0	0	0	1	0 – 58.3
Orchard Oriole (OO)	*Icterus spurius*	0	0	0	2	2	6.7 – 100
Brown-headed Cowbird (Cow)	*Molothrus ater*	11	1	8	12	32	N/A
Indigo Bunting (IB)	*Passerina cyanea*	0	0	0	1	1	0 – 71.4
Blue-gray Gnatcatcher (BG)	*Polioptila caerula*	0	0	2	0	2	0 – 75.9
Prothonotary Warbler (PW)	*Prothonotaria citrea*	0	0	1	1	2	6.7 – 20.9
Carolina Wren (CW)	*Thryothorus ludovicianus*	0	0	1	1	2	0 – 33
White-eyed Vireo (WV)	*Vireo griseus*	0	0	0	1	1	40
Hooded Warbler (HW)	*Setophaga citrina*	0	0	0	1	1	No data
Totals		12	1	14	21		

4.2.3. Clustering Analyses

Individuals were partitioned into five datasets in order to test the hypotheses outlined above: (A) ALL BIRDS, (B) JUVENILE COWBIRDS + ALL HOSTS, (C) ALL LOUISIANA BIRDS, (D) COWBIRDS ONLY, (E) HOSTS ONLY.

The microbial ecology package QIIME (Caporaso et al., 2010) was used for the following analyses. First, the reads were assigned to phylotypes at 97% sequence similarity because 3% is frequently cited as the "species" level of microbial taxonomy (Schloss & Handelsman, 2005). Next, we assigned taxonomies to OTUs using the RDP Classifier Program (Wang et al., 2007), with the default confidence threshold of 80%. A pairwise matrix of distances between each gut microbial community (i.e., each bird specimen) was constructed using UniFrac (Lozupone & Knight, 2005). UniFrac distances are calculated based on the amount of branch length in a phylogenetic tree that is unique to either of two environments (versus how much of the tree is shared by the environments). These distances can be based on presence-absence of OTUs or weighted by abundance. Our microbial phylogenetic tree was constructed with FastTree (Price et al., 2009). To reduce the effects of sampling (sequencing) bias, all individuals were randomly reduced to 5 018 reads, equal to the lowest number of reads for any bird in the dataset.

We constructed UPGMA dendrograms based on both the unweighted and weighted UniFrac distances to visually represent the relatedness of the gut microbiota for all five datasets and test the hypotheses. As a confidence metric, we jackknifed the trees using the QIIME recommendation of 75% of the reads used in the rarefaction (3 760) with 10 replicates. Principal coordinates analysis (PCoA) was also performed on both the weighted and unweighted UniFrac distance matrices.

As a complement to the phylogenetic-based methods, we visualized the data with nonmetric multidimensional scaling (NMDS). We square root-transformed the percentage of each sample that belonged to each bacterial phylum, then created a pairwise distance matrix using Bray-Curtis

dissimilarity, applied through the VEGDIST function of the VEGAN package (Oksanen et al., 2011) in R (R Development Core Team, 2010). The NMDS function of the ECODIST package (Goslee & Urban, 2007) was then used to calculate the two-dimensional positions of the samples (such that closer samples are more similar), the stress and R^2 value of the plot. Stress values >0.3 should not be considered valid whereas values <0.2 can be considered a good representation of the data (Quinn & Keough, 2002).

To specifically test the Convergence Hypothesis, we compared UniFrac distances between and within adult cowbirds, juvenile cowbirds and brood hosts. The assumption is that if adult cowbirds converge on a cowbird-specific microbiota, adults will have lower pairwise distances than within juveniles or either category to brood hosts. Both weighted and unweighted UniFrac distances were assessed.

4.2.4. Categorical Variable Significance

To look for a relationship between categorical variables (associated with each bird; Appendix D) and the microbial communities, we used the statistical tools Adonis (McArdle & Anderson, 2001) and Anosim (Clarke, 1993) implemented in QIIME. The categorical variables included family, genus, species, age (based on percent of skull ossification), locality (Louisiana or California), diet (mostly plant material, mostly animal material, both animal and plant material) and stomach contents (e.g., "insects" or "white millet"). We also tested the total number of phyla recorded per bird (bacterial richness) to see if the diversity of the established microbial community had an effect on the communities. We calculated significance of all variables for both the weighted and unweighted UniFrac distance matrices with 999 iterations; we also repeated analyses with a higher number of rarefied sequences for each dataset to see if the signal changed when more data points from fewer individuals were included. Datasets were rarefied to 5 018, 17 000 and 42 000 sequences.

Based on results, we ran another Adonis test to specifically partition the variation in the samples that was due to the age and locality variables. We used the ALL BIRDS dataset and analyzed the weighted UniFrac distance matrix, unweighted UniFrac distance matrix and a sites (birds) by species (bacterial phyla) matrix, where cells were assigned the value of the number of sequences belonging to each phylum for each bird. We used the ADONIS function of the VEGAN package in R and performed 999 iterations, constraining resampling to be within species. Since the order of variables being tested matters, we tried age then locality as well as locality then age for each of the datasets.

4.2.5. Comparison to Mammals and Insects

To put our results in a broader context, we compared the birds to a mammal dataset (Ley et al., 2008a) containing 56 individuals from 56 species across 13 orders (Appendix E) and an insect meta-analysis dataset containing 85 individuals from 62 species across seven orders (Colman et al., 2012). Although the mammal and insect datasets were collected with different methods than those outlined above, most sequence fragments contained the V6 region. We pruned all reads to the same length for analysis. We only analyzed samples with greater than 200 sequences. To increase coverage for some species we combined mammals belonging to the same species into single samples. This treatment should not skew the results of our analysis, because Ley et al.

(2008a) found that individuals from the same species clustered together. We taxonomically assigned reads using RDP Classifier Program. For PCoA, we rarefied all samples to 200 reads and used the unweighted UniFrac distances as input. We also performed NMDS on samples, as described above. To test for significant associations between class, order and diet categories (herbivore, carnivore, omnivore), we tested each variable against both the weighted and unweighted UniFrac distance matrices in the same manner as above.

4.3. RESULTS

Initial quality control steps resulted in 3 500 665 pairs of reads with no errors in priming sequence, region of overlap or individual tags. Three hundred and thirty three potentially chimeric sequences (0.01% of reads) and 62 201 sequences that did not align to the domain Bacteria (1.7% of reads) were removed. The reads passing these filters were included in the final dataset, totaling 3 438 131 sequences and averaging 70 165 sequences per individual, but reads/sample varied by two orders of magnitude (range: 5 018 - 629 093).

Four bacterial phyla were detected in all individuals: Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria. Proteobacteria and Firmicutes dominated most of the samples (Fig. 4.2). Proteobacteria constituted an average of 54.7% of sequence reads for an individual, Firmicutes an average of 36.0%, and Actinobacteria and Bacteroidetes an average of 1.3% and 1.7%, respectively. An additional 16 phyla were identified: Acidobacteria, Chloroflexi, Cyanobacteria, Deinococcus-Thermus, Fusobacteria, Gemmatimonadetes, Nitrospira, OD1, OP10, OP11, Planctomycetes, Spirochaetes, TM7, Tenericutes, Thermotogae, Verrucomicrobia. 5.8% of sequences were from unknown phyla within Bacteria. All birds shared 36 genera (Table 4.2) out of 445 (8%) identified; an additional 139 genus-level OTUs did not align to known genera. Cowbirds harbored a higher average number of bacterial phyla (12.72) than brood hosts (12.59) but this difference was not significant (one tailed t-test, p=0.40).

4.3.1. Clustering Analyses

Cowbird samples did not cluster together in the UPGMA dendrogram of ALL BIRDS (Fig. 4.3). Brood host species having more than one individual also did not cluster together, even when cowbird samples were excluded from the analyses (i.e., JUST HOSTS dataset, Fig. 4.4). NMDS representation of ALL BIRDS showed little segregation by age or locality (Fig. 4.3), although the stress of the plot was low (0.1395). In general, UPGMA dendrograms and NMDS plots of all datasets showed little clustering by bird species and high levels of variation (Fig. 4.4). The two replicate samples, NorthernCardinal4.1 and NorthernCardinal4.2 were always most closely related to each other in every analysis.

Pairwise distances were assessed between and within adult cowbirds, juvenile cowbirds and brood hosts to test whether adult cowbirds converged on a cowbird specific microbiota (i.e., adult cowbird microbiota were more similar to each other than they were to other group comparisons or other groups were to themselves). All pairwise comparisons between and within groups had largely overlapping distributions (Fig. 4.5) for both weighted and unweighted UniFrac distances, and thus, adults were not more similar to each other than other comparisons.

Table 4.2. The 36 bacterial genera shared by all birds, assigned by RDP Classifier.

Phylum	Class	Order	Family	Genus
Actinobacteria	Actinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium
Actinobacteria	Actinobacteria	Actinomycetales	Dietziaceae	Dietzia
Actinobacteria	Actinobacteria	Actinomycetales	Micrococcaceae	Micrococcus
Actinobacteria	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium
Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Chryseobacterium
Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Planobacterium
Firmicutes	Bacilli	Bacillales	Bacillaceae	Anoxybacillus
Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus
Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Marinilactibacillus
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Leuconostoc
Firmicutes	Bacilli	Lactobacillales	Leuconostocaceae	Weissella
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium
Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Veillonella
Proteobacteria	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Methylobacterium
Proteobacteria	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Xanthobacter
Proteobacteria	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingobium
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiales incertae sedis	Aquabacterium
Proteobacteria	Betaproteobacteria	Burkholderiales	Burkholderiales incertae sedis	Tepidimonas
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Acidovorax
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Curvibacter
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Diaphorobacter
Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	Schlegelella
Proteobacteria	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Janthinobacterium
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Campylobacteraceae	Campylobacter
Proteobacteria	Epsilonproteobacteria	Campylobacterales	Helicobacteraceae	Helicobacter
Proteobacteria	Gammaproteobacteria	Aeromonadales	Aeromonadaceae	Aeromonas
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Cronobacter
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Kluyvera
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter
Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas
Proteobacteria	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Stenotrophomonas

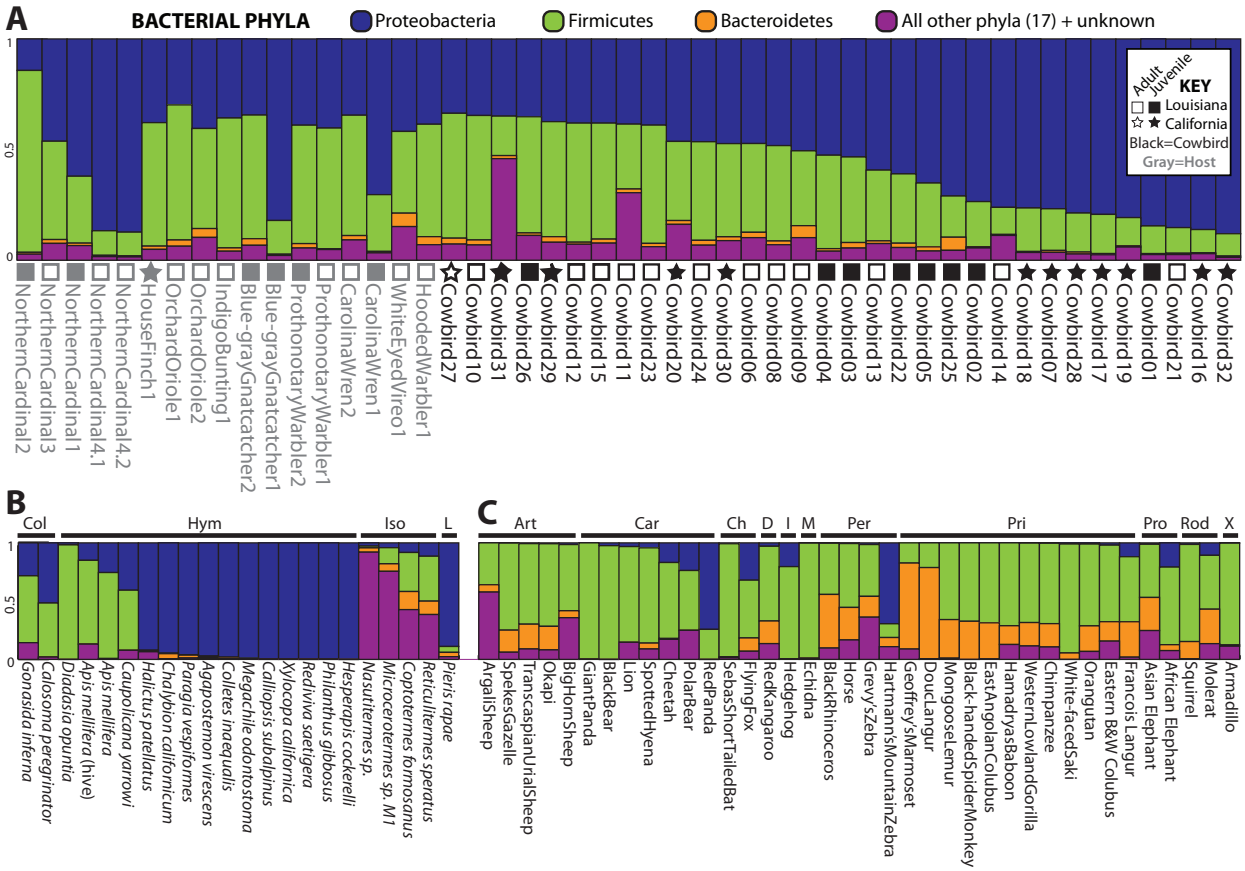


Figure 4.2. Relative abundance of the top three bacterial phyla in each sample for (A) birds, (B) insects and (C) mammals with greater than 200 reads. Locality and adult/juvenile status are shown for each bird; cowbirds are labeled with black text and symbols, brood hosts with gray. Insect and mammal orders are depicted by bars across the top of their graphs. Insect orders: COLeoptera, HYMenoptera, ISOptera, Lepidoptera. Mammal orders: ARTiodactyla, CARnivora, CHiroptera, Diprotodontia, Insectivora, Monotremata, PERissodactyla, PRImates, PROboscidea, RODentia, Xenarthra.

4.3.2. Categorical Variable Significance

A total of 12 statistical tests was computed for each categorical variable for each dataset (Fig. 4.6E). The locality category was significant in eight of the 12 tests for the three datasets: ALL BIRDS, JUVENILE COWBIRDS + ALL HOSTS and COWBIRDS ONLY (Fig. 4.6). Age was significant in seven of the 12 tests in the ALL BIRDS datasets, in five of 12 tests for ALL LOUISIANA BIRDS and in three of 12 tests for COWBIRDS ONLY. The taxonomic categories (bird family, genus and species) were not significant for HOSTS ONLY, were significant in one or two tests for ALL BIRDS and JUVENILE COWBIRDS + HOSTS and were significant in five of 12 tests for ALL LOUISIANA BIRDS. Diet was significant in four of the tests for the ALL BIRDS and JUVENILE COWBIRDS + ALL HOSTS datasets and in nine tests for ALL LOUISIANA BIRDS. Stomach contents were only significant in one test; bacterial richness was never significant.

To further investigate two frequently significant variables, locality and age, we conducted three Adonis tests to specifically partition the variance attributable to each variable. We ran each test

twice, varying the order of the variables, since this can have an affect on the results, and used three distance matrices. Age was significant in all six tests and locality was significant in one of six tests (Table 4.2). An age-by-locality interaction was not significant.

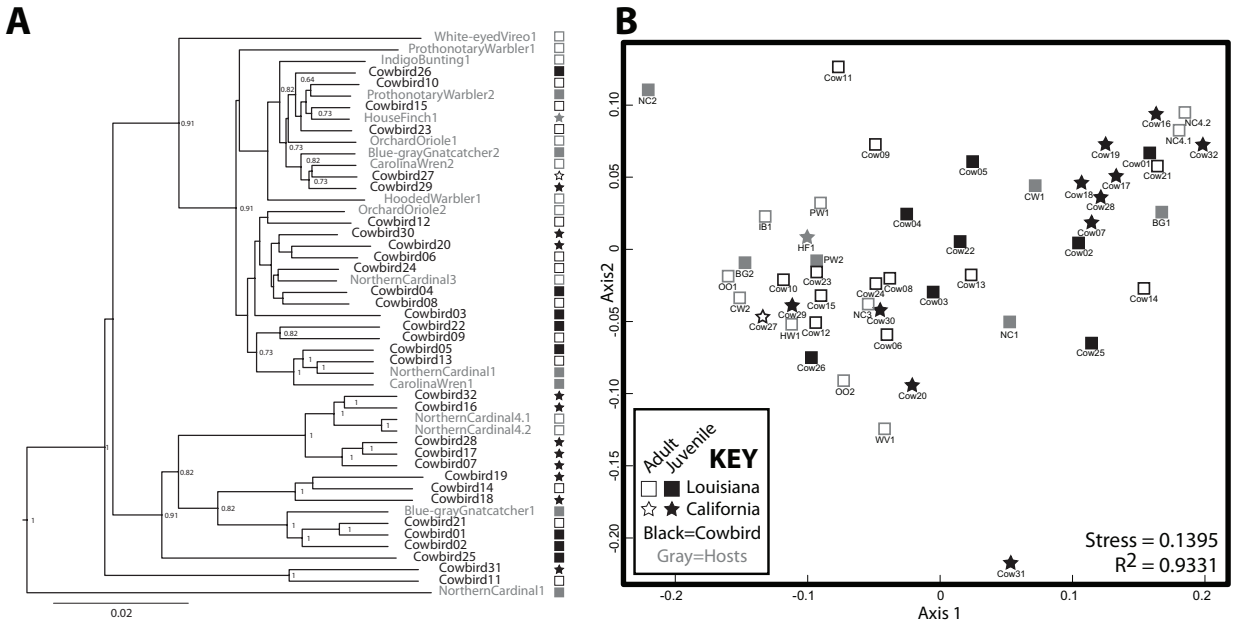


Figure 4.3. (A) Dendrogram of gut microbiota relatedness based on weighted UniFrac distances; all samples rarefied to 5 018 reads; jackknifed support values are shown for nodes where support >0.70. (B) NMDS ordination of Bray-Curtis dissimilarities of microbiota composition (see section 4.2. Methods). Cowbirds are labeled with black text and symbols, hosts with gray.

4.3.3. Comparison to Mammals and Insects

For the taxonomic assignments and NMDS analyses, we only included individuals with more than 200 sequences. For the PCoA, we rarefied all individuals to 200 sequences and samples with less than 200 sequences were not used. This reduced the mammal dataset to 38 samples belonging to 11 orders and the insect dataset to 22 individuals belonging to four orders (Appendix E). Consistent with Ley et al. (2008a), mammals were dominated by Firmicutes and Bacteroidetes (Fig. 4.2), whereas bird samples were predominately Firmicutes, with some samples having mostly Proteobacteria or Bacteroidetes (Fig. 4.2). Insects also generally contained a majority of Firmicutes although individual samples varied between 100% Proteobacteria and 98% Firmicutes (Fig. 4.2). The PCoA showed birds as distinct from mammals and insects, which largely clustered into their respective groups but contained some overlap (Fig. 4.5); birds spanned a greater portion of PC2 than all Mammals. NMDS was broadly overlapping but birds clustered together in the middle of the plot (Fig. 4.7). Independent Adonis tests for significance of class, order and diet categories revealed highly significant associations (p<0.01) between the gut microbiota and all three variables (Table 4.3), except for diet and the weighted UniFrac distance matrix, which was not significant (p=0.206).

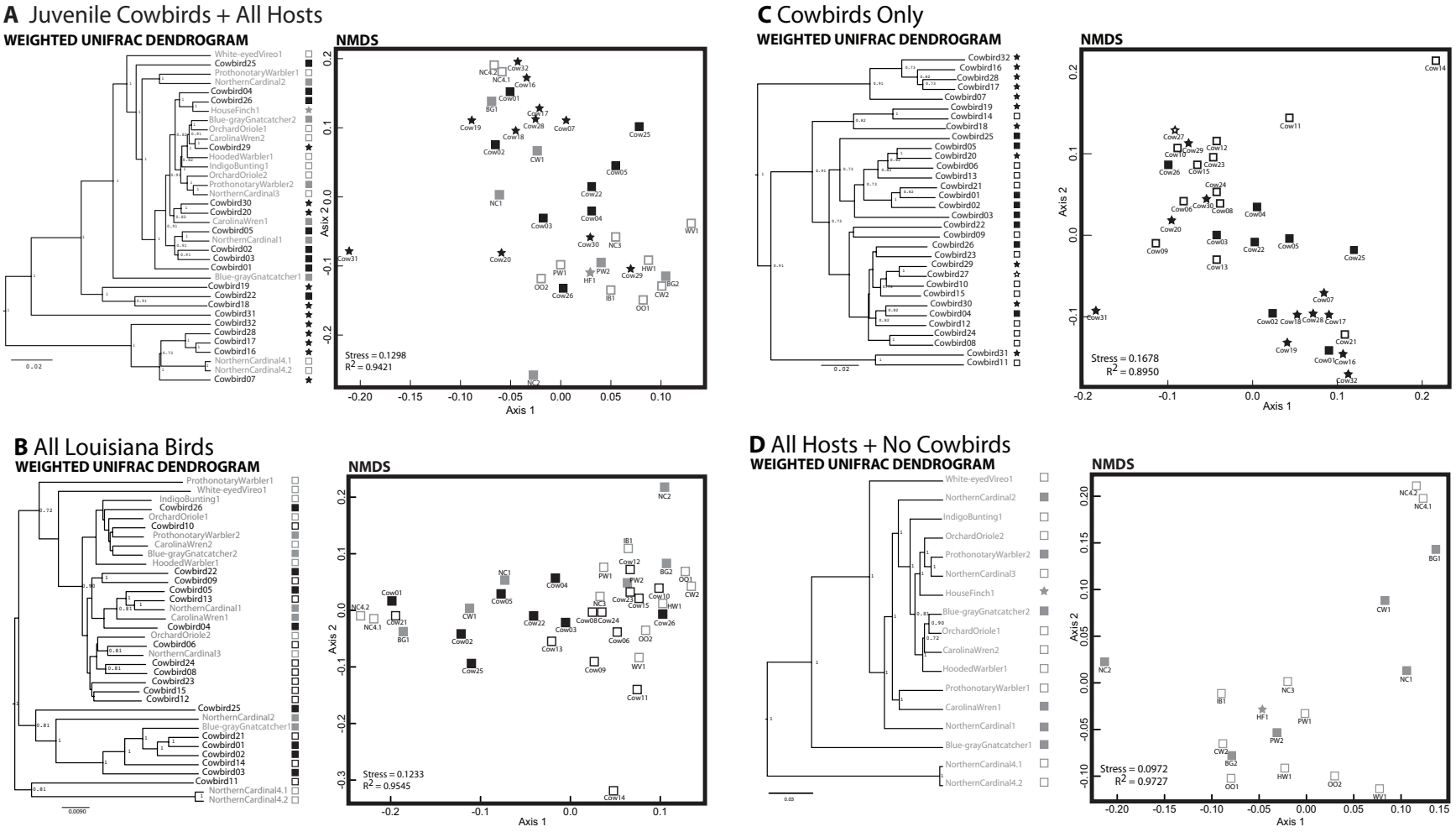


Figure 4.4. Dendrogram and NMDS for (A) Juvenile Cowbirds + All Hosts, (B) All Louisiana Birds, (C) Cowbirds Only and (D) Hosts Only datasets.

Table 4.3. R^2 values of Adonis test for significance of locality and age across the weighted and unweighted UniFrac distance matrices and the raw sites by species (birds by bacterial phyla) matrix. Asterisks indicate *p*-values. Since the order of variables matters, tests were conducted where each variable was ordered first.

	Weighted	Unweighted	Counts		Weighted	Unweighted	Counts
Locality	**0.066****	0.016	0.027	Age	**0.249*****	**0.115***	**0.169****
Age	**0.212*****	**0.116***	**0.147***	Locality	0.029	0.016	0.005
Locality:Age	0.008	0.014	0.008	Age:Locality	0.008	0.014	0.008

***<0.01, **<0.05, *<0.10

Table 4.4. R^2 values of independent Adonis tests for significance of class (mammal, insect, bird), taxonomic order and diet (HCO; herbivore, omnivore, carnivore) for the weighted and unweighted UniFrac distance matrices. Asterisks indicate *p*-values.

	Weighted	Unweighted
Class	**0.139*****	**0.418*****
Order	**0.897*****	**0.544*****
HCO	0.029	**0.080*****

***<0.01, **<0.05, *<0.10

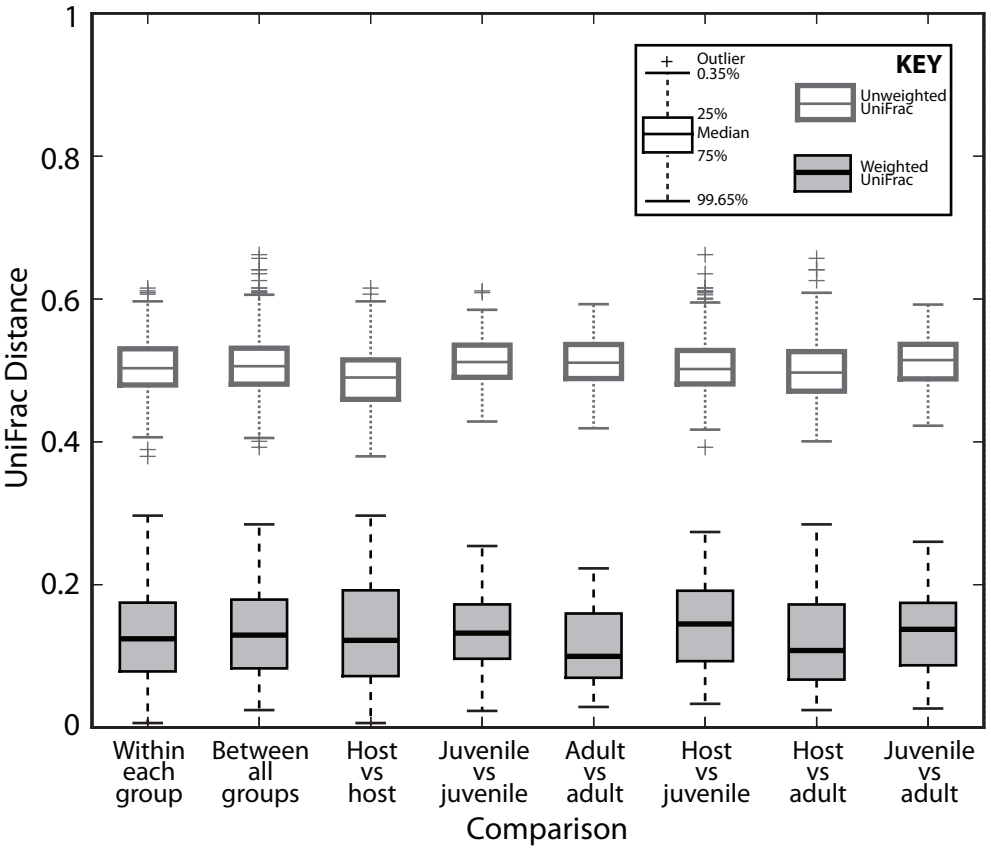


Figure 4.5. Box plot of UniFrac distances within and between hosts, adult cowbirds and juvenile cowbirds.

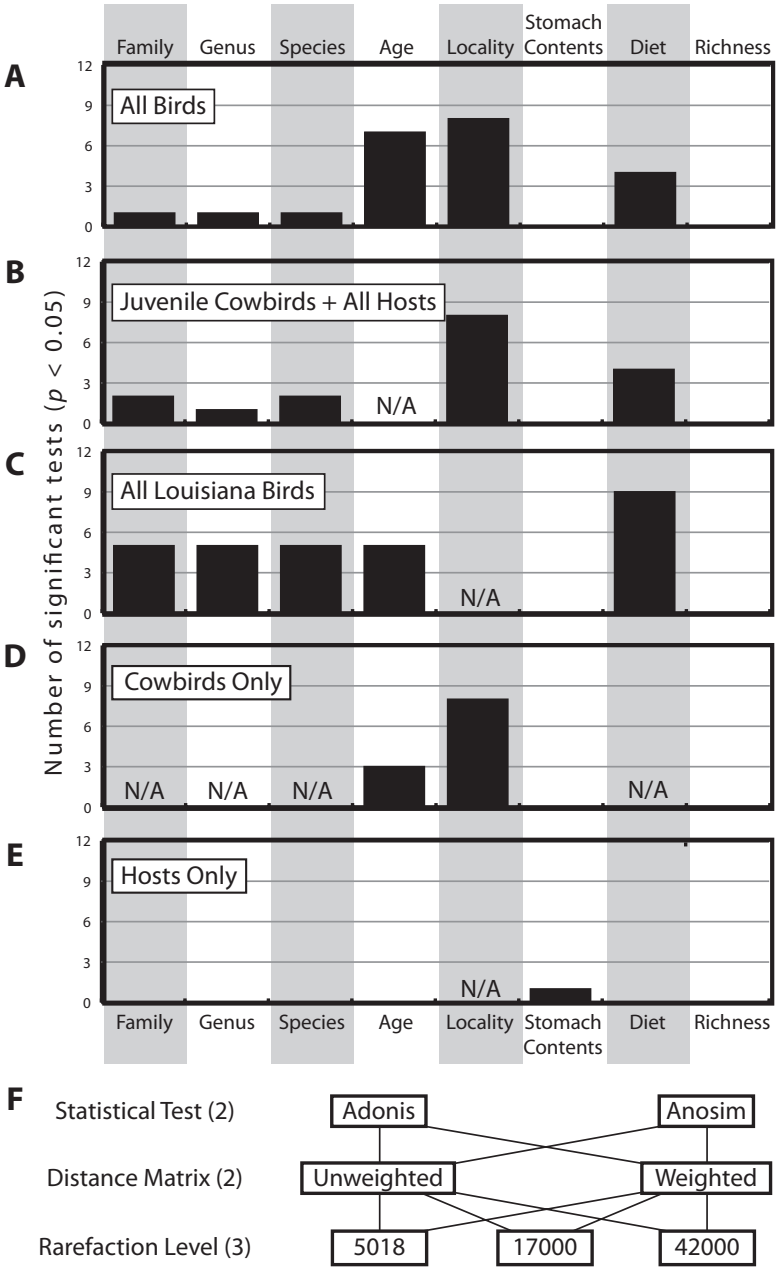


Figure 4.6. Histogram of how many times each of the categorical variables was significant at the $p<0.05$ level (out of 12 total tests, see setion 4.2. Methods), for (A) All Birds, (B) Juvenile Cowbirds + All Hosts, (C) All Louisiana Birds, (D) Just Cowbirds and (E) Just Hosts. Note some categorical variables were not applicable to some datasets (due to one or fewer individuals belonging to one or more categories).

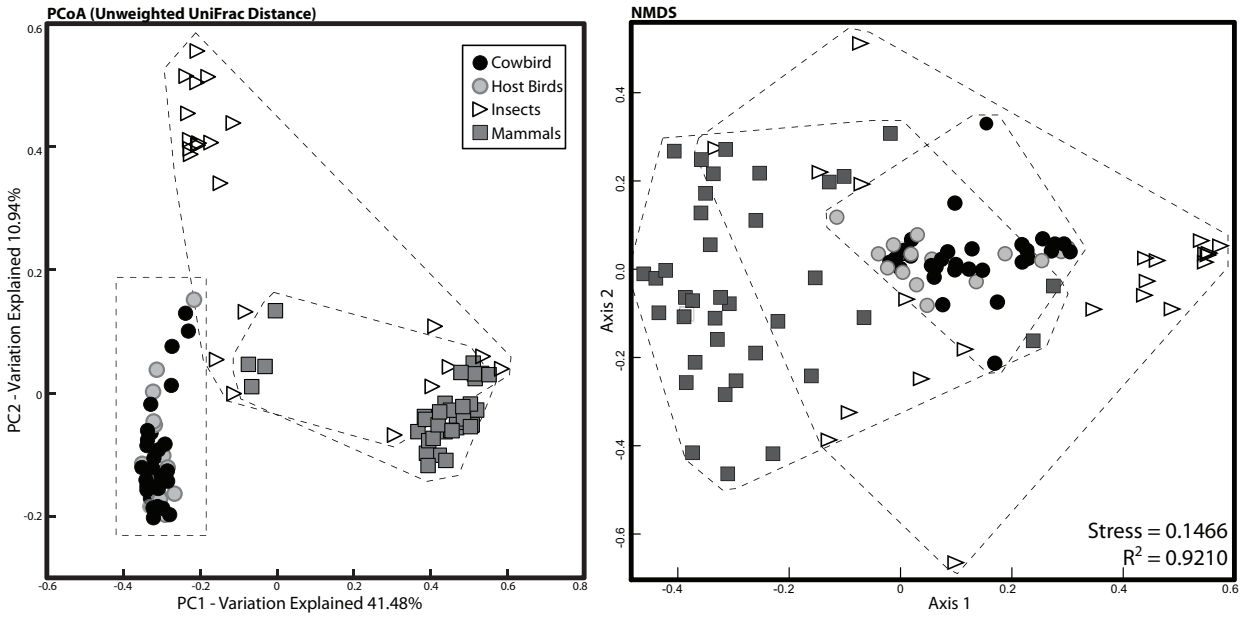


Figure 4.7. Relationship between mammal, insect and bird microbiota. (A) Principal coordinates analysis of unweighted UniFrac distances. (B) Nonmetric multidimensional scaling, using relative abundance of bacterial phyla as input (see Methods). Birds are denoted by circles (cowbirds are black, brood hosts are gray), mammals are squares and insects are triangles. Dashed lines encapsulate all birds, all insects and all mammals.

4.4. DISCUSSION

The microorganisms living on and in vertebrates provide essential functions for the host, yet how these complex communities arise and remain stable is largely unknown. Many factors seem important, including intrinsic (e.g., genetics and physiology) and extrinsic (e.g., ecology and the environment) components. Cowbirds provide a natural system to study these components where vertical transmission of microbiota through genetics is decoupled from horizontal transmission through parental care. Our results indicate that within cowbirds, the local environment is most significantly correlated with microbiota similarity (Fig. 4.4), whereas bird taxonomy, contents of the stomach and bacterial richness of the gut were not. Thus, of our four hypotheses (Fig. 4.1), the Environment Hypothesis is the best supported. Although two obvious Louisiana and California clusters were not observed in any analysis, some locality-specific groups were identified – e.g., five juvenile cowbirds from California always grouped together (Cowbirds 7, 16, 17, 28, 32; Figs. 4.3, 4.4). Detecting a difference between these two localities is not terribly surprising; population level differentiation in gut microbiota has been detected in other organisms, including birds (Godoy-Vitorino et al., 2012).

It is important to note that we cannot ascribe the differences between California and Louisiana cowbirds to specific factors, since many aspects of the environment are captured within the "Locality" variable. Geographic distance is positively correlated with microbiota dissimilarity (Dominguez-Bello & Blaser, 2011), but local flora and fauna, photoperiod, available food, weather conditions, etc. may all affect microbiota and differ by locality. Further experimentation including gathering extensive environmental data on samples that share some of the potentially important factors but that differ on others may add resolution. Environmental niche modeling on

microbes/microbiota may indicate what abiotic environmental parameters are most important for shaping host-associated microbiota.

Diet and age were the next most frequently significant variables. Dietary specialization is an important contributor to mammal (Ley et al., 2008b) and insect (Colman et al., 2012) gut microbiotas and our results indicate this may also be true for birds (Figs. 4.6, 4.8, Table 4.3), although stomach contents was not a significant predictor variable in any of our analyses. How are these data compatible? In mammals, dietary specialization drives convergence in gut communities (Muegge et al., 2011) and once the community has stabilized, it is relatively immune to perturbation (Walter & Ley, 2011). This coincides very well with the apparent importance of diet but not of actual stomach contents in cowbirds. Diet may be as important as locality in shaping the microbiota – when just Louisiana birds were analyzed, the diet variable became significant in 75% of the tests. Another thing to note is that many birds had white millet in their stomachs; how human-supplied bird food (e.g., non-native seeds) impacts bird gut microbiota is an interesting question and worth investigation.

Very young animals frequently have a distinct gut microbiota from adults and undergo a transition period before reaching a stable, adult-like community (Vaishampayan et al., 2010). Although the significance of the age variable (Fig. 4.4) implies differentiation between younger and older individuals, adult cowbirds were no more similar to each other than they were to juveniles or hosts (Fig. 4.5) and they never formed their own cluster to the exclusion of juveniles and brood hosts in any analysis. Therefore, we reject the Convergence Hypothesis, but leave room for a transition occurring between juvenile and adult cowbirds. One difficult but ideal experiment to assess the relationship between age and brood parasite gut microbiota would be to track individuals through time. Resampling the same individuals at successive time points would monitor the process of gut maturation and allow comparisons across species and brood parasite/brood host.

High inter-individual variation appears to be a hallmark of microbiota studies, so much so that the concept of a "core microbiota" is in doubt (Lozupone et al., 2012). The birds in this study belong to a single order, the Passeriformes, and appear to have more variation than any of the non-avian orders we analyzed, with the exception of Hymentoptera (Fig. 4.8). Cowbirds in particular appear to have a highly variable gut microbiota. The relative contribution of Proteobacteria and Firmicutes to cowbirds spans nearly all brood host species and the cowbirds span nearly the entire dendrogram and NMDS plot (Fig. 4.3). It is unclear whether this is an adaptation to a generalist brood parasitic lifestyle. It makes intuitive sense that being capable of utilizing a large taxonomic breadth of microbial symbionts would be useful for a generalist, especially if early environment shapes gut microbiota. However, we did not find that cowbirds had significantly higher richness than brood hosts. Additionally, the four Northern Cardinal samples appear to contain as much variation as the cowbirds (Figs. 4.2, 4.3). Importantly, the two replicates from the same extraction, NorthernCardinal4, always grouped together, though, which we take as an indication that this is not a methodological error.

Despite frequent high levels of variation, taxonomic signal of the vertebrate is frequently detected in microbiota studies. The bird, mammal and insect classes were significantly different (Table 4.3), but at lower taxonomic levels within the birds, we found minimal associations with the gut microbiota (Fig. 4.4). Across analyses and datasets, cowbirds and hosts were interspersed

(Fig. 4.3) – neither cowbirds nor individuals from each brood host species clustered together. We therefore reject the Nature Hypothesis since there is no evidence for a specific cowbird gut community shared by all cowbirds. We are also unable to appropriately evaluate the Nurture Hypothesis, since an underlying assumption was that brood host species would be most similar to one another. The high variation in the system and the importance of locality and diet are consistent with an important role of ecology in shaping gut microbiota, so the Nuture Hypothesis remains a viable explanation. For example, if the factors that shape gut microbiota are drawn from largely overlapping ecological niche space, a lack of bird taxonomic signal and generally low levels of clustering would be expected, in addition to the importance of parameters like locality and diet. An ideal experiment to test the Nurture Hypothesis would be to sample entire brood families from a single nest (parents, offspring, brood parasites) plus the cowbird mother. Under this framework, there would be two sets of parents and offspring (cowbird and brood host) and known brood host species and nest-mates. Furthermore, swabs of a nest could be taken (to directly compare the nest bacterial composition). This experiment would be logistically difficult to accomplish but would allow direct comparisons between juvenile cowbird microbiota, its genetic mother, the brood host parents and siblings and the immediate environment.

Another caveat with the Nurture Hypothesis comes from the sampling, which included only samples the LSUMNS had previously collected. While some juveniles are most closely related to brood host individuals, we have no way of evaluating whether that is because they were raised by that species. The brown-headed cowbird parasitizes over 200 bird species, only nine of which are represented here and of those sampled, there is an uneven distribution across age and locality (Table 4.1). This sampling is not ideal, but using museum specimens in this way represents information gained from a single sampling effort. Based on these results, targeted future studies may want to include as much taxonomic breadth as possible, with replicates from each locality. It was not anticipated that brood hosts would not cluster together, since they are genetically, ecologically and presumably environmentally most similar. We recommend including more than one individual from each species, when possible.

This study was conducted using a single marker and it relies on OTUs delimited from these genetic data. Metagenomic studies randomly sequence many loci across a microbiota sample and show that while taxonomic identity can vary widely across individuals, functional groups are highly conserved (Lozupone et al., 2012). The overlapping ecology, lack of taxonomic signal and significant effect of sampling locality indicate this would be an interesting application of metagenomics. Does the brood parasite's gut contain more functional categories than a traditional bird? What, if any, functional categories are most represented? Retesting all of the above hypotheses with metagenomic data instead of fingerprint data would be valuable.

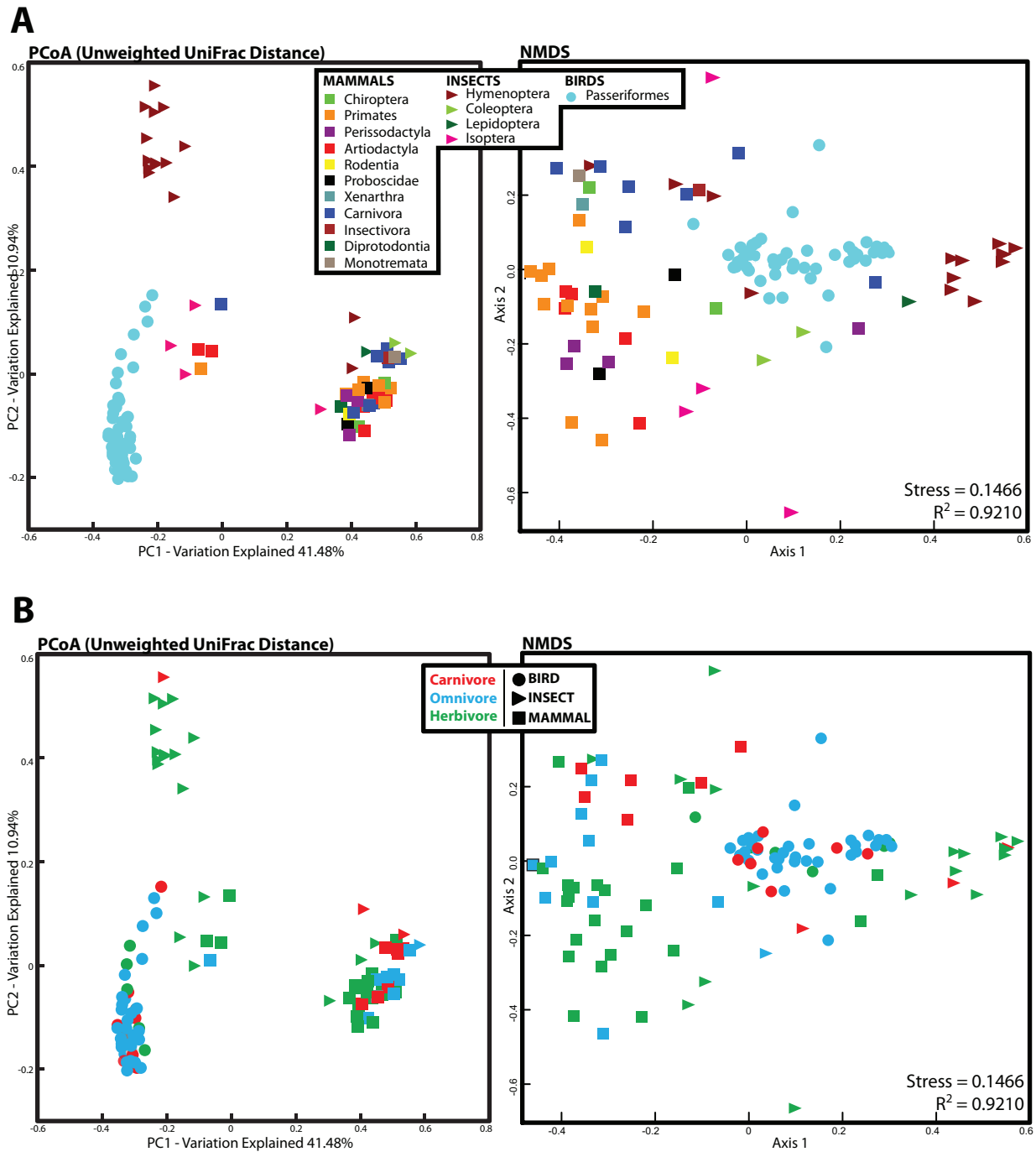


Figure 4.8. Relationship between mammal, insect and bird microbiota shown with principal coordinates analysis of unweighted UniFrac distances (left column) and nonmetric multidimensional scaling, using relative abundance of bacterial phyla as input (right column, see section 4.2. Methods). Results are the same as Figure 5 from the original manuscript, just colored by different attributes. Birds are denoted by circles (cowbirds are black, brood hosts are gray), mammals are squares and insects are triangles. Dashed lines encapsulate all birds, all insects and all mammals. Top Row: Colors denote taxonomic order of the vertebrate. Bottom Row: Colors denote dietary specialization: Red is carnivorous/mostly animal material. Green is herbivorous/mostly plant material. Blue is omnivorous/mix of plant and animal material.

Chapter 5.
Assessing The Use Of Gut Microbiota As A Marker For Phylogeographic Inference In Neotropical Birds

5.1. INTRODUCTION

Phylogeography aims to understand the arrangement of genetic lineages across space and through time. Through phylogeographic inference, we learn about evolutionary and biogeographic processes that have shaped the biodiversity around us. Multi-locus data improve these inferences (Brito & Edwards, 2009) and statistical analyses (such as hypothesis testing) provide a rigorous framework for incorporating multiple, potentially conflicting, markers (Knowles & Maddison, 2002). With the rise of high-throughput sequencing, the addition of loci to a phylogeography project has become both cheaper and easier (Glenn, 2011, McCormack et al., 2013). A parallel rise in culture independent methods in microbial ecology has facilitated the genetic cataloguing, description and analysis of complex microbial communities (Frank & Pace, 2008). Research regarding the microorganisms that live on and in larger host organisms (the microbiota) has flourished using these new technologies and we now know vertebrates house trillions of microbes, most of which are neutral or beneficial to the health of the host (Sears, 2005). The gut microbiota is one of the most densely populated natural environments ever described (Whitman et al., 1998), comprising up to 40,000 species (Xu & Gordon, 2003, Frank & Pace, 2008) across Bacteria, Archaea, Eukarya and viruses (Rajilić‐Stojanović et al., 2007, Frank & Pace, 2008).

There are ample reasons to believe the gut microbiota contain novel evolutionary information about the host. First, these communities are vital for normal physiological function of the host (Qin et al., 2010) and together host and gut microbiota form an obligate symbiosis (Xu & Gordon, 2003). Microbes and host are under many levels of shared and independent selection and the fitness of each is dependent on the other (Ley et al., 2006). Many obligate interactions between disparate taxa result in concordant evolutionary history and result in novel biological insight about the organisms involved, for example, yuccas and yucca moths (Yoder et al., 2010), pocket gophers and chewing lice (Hafner & Nadler, 1988), hawks and ectoparasites (Whiteman et al., 2007), humans and HIV (Gao et al., 1999), etc.

Second, host genotype has been shown to directly affect gut communities through transplant experiments (Rawls et al., 2006), model organisms (Fraune & Bosch, 2007, Ley et al., 2005) and twin studies (Van de Merwe et al., 1983, Stewart et al., 2005, Zoetendal et al., 2001). Additionally, phylogenetic relationships among hosts have been recovered at multiple taxonomic levels (Ochman et al., 2010, Phillips et al., 2012, Ley et al., 2008a). The dual effects of host genetics and phylogenetics imply the gut microbiota may reflect evolutionary history of the host.

Third, spatial variation – a necessity for phylogeographic investigation – can be recovered by gut microbiota (Hird et al., In Review, Koopman & Carstens, 2011, Godoy-Vitorino et al., 2012). Individual microbial species can be used to track migration of a host (Dominguez-Bello & Blaser, 2011) and human commensal bacteria have corroborated specific dispersal events (Moodley et al., 2009).

Finally, some aspects of the gut microbiota can capture ecology of the host. Most notably, long-term dietary specialization has caused gut microbiota of phylogenetically distant mammals to converge on similar communities (Muegge et al., 2011). Modification of diet in the short term (1 generation) causes gut microbiota to diverge in many groups, including birds (Wienemann et al., 2011, Blanco et al., 2006). The gut microbiota can also influence behavior in a multitude of ways (Ezenwa et al., 2012, Heijtz et al., 2011), including causing assortative mating in flies (Sharon et al., 2010).

Neotropical birds and the complex geologic history of Central and South America provide the phylogeographic framework with which to investigate the various signals of phylogeographic interest to contributing factors to avian microbiota. Phylogeography is traditionally interested in the geographic distribution of gene lineages within a species and comparative phylogeography compares gene lineages across species; incorporating microbial markers into phylogeography expands the scale of investigation and may inform about both host and microbes. Here, we sampled taxonomically and ecologically diverse birds from across the central mountain ranges of Costa Rica (Fig. 5.1) in order to (1) catalogue bacterial diversity of Neotropical bird guts, (2) quantify the contributions of host genotype, ecology, spatial distribution and microbial community to gut microbiota diversity and (3) assess signals of evolutionary history within the gut. Finally, we discuss the utility of gut microbiota as a marker in phylogeography.

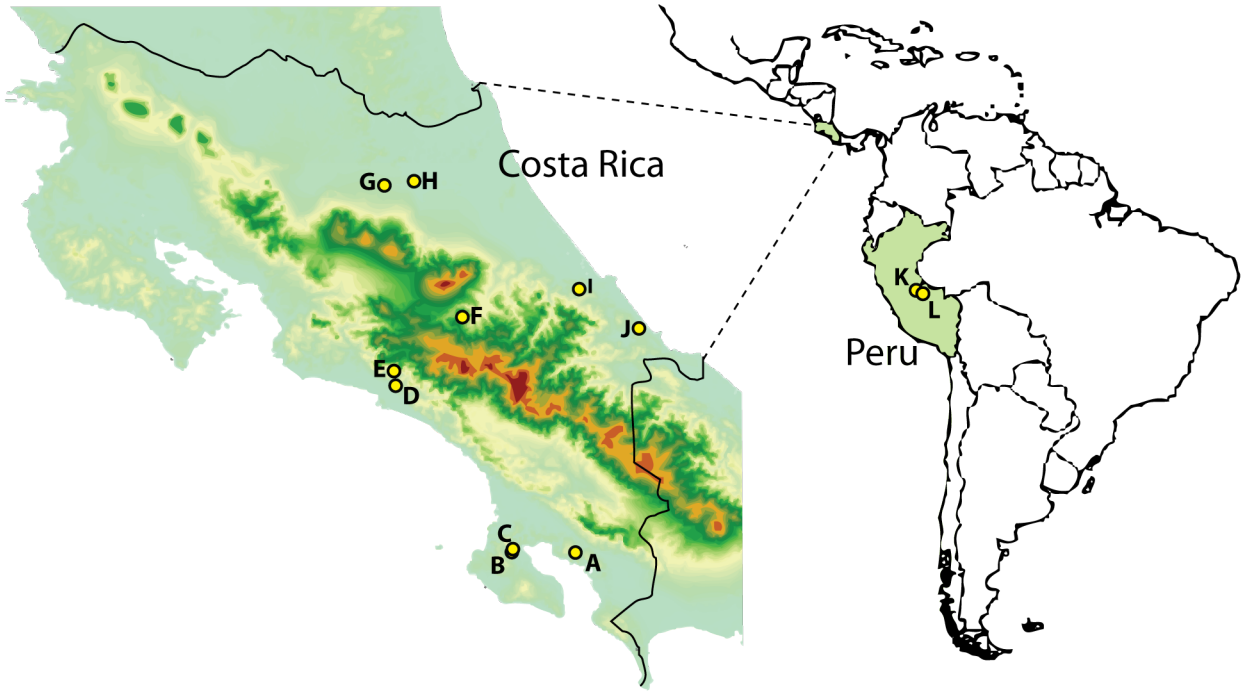


Figure 5.1. Sampling localities in Costa Rica (left) and Peru (right). A: Piedras Blancas, B: Los Charcos, C: Golfo Dulce, D: Londres, E: Santa Juana, F: El Copal, G: Tirimbina, H: La Selva, I: Veragua, J: Tuba Creek, K: Janirvan, L: San Jorge II. Warmer colors indicate higher elevations.

5.2. MATERIALS AND METHODS

5.2.1. Sampling

During fieldwork conducted between May and August 2010, the large intestine was extracted from 108 birds in Costa Rica and seven in Peru (Table 5.1, Fig. 5.1, Appendix F). They were immediately stored in liquid nitrogen, following the protocol of Godoy-Vitorino et al. (2010) and kept frozen until DNA extraction at LSUMNS molecular lab. We focused on bird species whose ranges span the mountains of Costa Rica and extend south into Peru.

Following Gloor et al. (2010) we used combinatoric primers and massive multiplexing of PCR amplicons for sequencing on an Illumina Hi-Seq. This method uses paired-end sequencing technology to generate pairs of sequences with 100% overlap across variable region 6 (V6) of the 16S component of rRNA; primer sequences align to positions 967-985 and 1078-1061 on *Escherichia coli* 16S rRNA (Gloor et al., 2010). We chose the V6 region of 16S for our fingerprint marker because of its ubiquity in bacteria, its ease of universal amplification (primers can flank highly variable regions by placing them in highly conserved regions), appropriate level of variability for our question and relative lack of horizontal gene transfer (Clarridge III, 2004). One bird was amplified and sequenced a second time from a single extraction; these replicates are *Cyanocompsa.cyanoides.1.1* and *Cyanocompsa.cyanoides.1.2.* Three birds had two extractions completed and were sequenced independently: *Attila.spadiceus.1, Trogon.rufus.2* and *Nyctidromus.albicollis.1. Attila spadiceus* had one extraction from the posterior large intestine and the other from the anterior large intestine; *T. rufus* had two extractions in tandem from the posterior large intestine; *N. albicollis* had one extraction from the posterior large intestine and a second from one of the ceca. These replicates were intended to quantify differences along the digestive tract and/or how sensitive the methods are to these differences.

We used several measures of sequence quality control. First, both reads of a given pair had to match across 100% of the bases. The pairs also must have no errors in the individual tag or priming sequence. We used the BELLEROPHON (Huber et al., 2004) function within the MOTHUR program (Schloss et al., 2009) to identify and discard potentially chimeric sequences. Finally, we used MOTHUR to discard sequences that did not blast to the domain Bacteria. The reads passing these filters were included in the final dataset.

5.2.2. Subsampled Datasets

To assess patterns across different spatial, taxonomic and ecological scales, we subdivided the dataset eight ways.

1. FULL DATASET: all samples. (N=116)
2. PASSERINES: all the birds belonging to the order Passeriformes. (N=88)
3. NON-PASSERINES: all the birds belonging to orders other than Passeriformes. (N=27)
4. > 2 INDIVIDUALS: all individuals belonging to species sampled more than once. This allowed us to look for clustering within species. (N=80)
5. *CYANOIDES*: all individuals belonging to the species *Cyanoides cyanocompsa*. This removed all taxonomic variation from the dataset. (N=8)

Table 5.1. Order, family, genus, species, sampling locality and number of samples used in this study; sampling localities mapped on Figure 5.1.

				Sampling Locality										
Order	Family	Genus	Species	A	B	C	D	E	F	G	H	I	J	K	L
Apodiformes	Trochilidae	*Amazilia*	*tzacatl*							1					
Apodiformes	Trochilidae	*Florisuga*	*mellivora*			2				1					1
Apodiformes	Trochilidae	*Phaethornis*	*longirostris*		2	1									
Apodiformes	Trochilidae	*Thalurania*	*colombica*			2									
Apodiformes	Trochilidae	*Threnetes*	*ruckeri*		1					3					
Caprimulgiformes	Caprimulgidae	*Nyctidromus*	*albicollis*									2			
Columbiformes	Columbidae	*Geotrygon*	*montana*			1									
Coraciiformes	Momotidae	*Baryphthengus*	*martii*							1					
Cuculiformes	Cuculidae	*Piaya*	*cayana*									1			
Passeriformes	Cardinalidae	*Cyanocompsa*	*cyanoides*			2				2				3	1
Passeriformes	Cardinalidae	*Habia*	*atrimaxillaris*		1	1									
Passeriformes	Cardinalidae	*Habia*	*fuscicauda*							4					
Passeriformes	Emberizidae	*Arremon*	*aurantiirostris*				1			1					
Passeriformes	Emberizidae	*Arremonops*	*conirostris*			1									
Passeriformes	Formicariidae	*Formicarius*	*analis*		1		1								
Passeriformes	Furnariidae	*Automolus*	*ochrolaemus*	1											
Passeriformes	Furnariidae	*Dendrocincla*	*fuliginosa*							1					
Passeriformes	Furnariidae	*Glyphorhynchus*	*spirurus*								1	1			
Passeriformes	Furnariidae	*Xiphorhynchus*	*susurrans*			2				1					
Passeriformes	Icteridae	*Cacicus*	*uropygialis*			2			1						
Passeriformes	IncertaeSedis	*Saltator*	*maximus*							1					
Passeriformes	Parulidae	*Myiothlypis*	*fulvicauda*					1							
Passeriformes	Pipridae	*Manacus*	*aurantiacus*			1									
Passeriformes	Pipridae	*Manacus*	*candei*							6					
Passeriformes	Pipridae	*Pipra*	*mentalis*		1	2	1			3					
Passeriformes	Thamnophilidae	*Cymbilaimus*	*lineatus*										1		
Passeriformes	Thamnophilidae	*Gymnopithys*	*leucaspis*			2				1					
Passeriformes	Thamnophilidae	*Hylophylax*	*naevioides*							1					
Passeriformes	Thamnophilidae	*Microrhopias*	*quixensis*			1			1						
Passeriformes	Thamnophilidae	*Myrmeciza*	*exsul*							1			1		
Passeriformes	Thraupidae	*Chlorophanes*	*spiza*			1									
Passeriformes	Thraupidae	*Oryzoborus*	*funereus*							1					
Passeriformes	Thraupidae	*Ramphocelus*	*costaricensis*			1									
Passeriformes	Thraupidae	*Ramphocelus*	*passerinii*							3					
Passeriformes	Thraupidae	*Sporophila*	*corvina*							1					
Passeriformes	Thraupidae	*Tachyphonus*	*luctuosus*									1			
Passeriformes	Thraupidae	*Tangara*	*gyrola*			1									
Passeriformes	Thraupidae	*Tangara*	*larvata*							1		1			
Passeriformes	Thraupidae	*Thraupis*	*episcopus*							1					

Table 5.1. Continued.

				Sampling Locality											
Order	Family	Genus	Species	A	B	C	D	E	F	G	H	I	J	K	L
Passeriformes	Thraupidae	*Volatinia*	*jacarina*							1					
Passeriformes	Tityridae	*Tityra*	*inquisitor*			1									
Passeriformes	Troglodytidae	*Cantorchilus*	*nigricapillus*									1			
Passeriformes	Troglodytidae	*Henicorhina*	*leucosticta*							1		1			
Passeriformes	Turdidae	*Turdus*	*grayi*							1					
Passeriformes	Tyrannidae	*Attila*	*spadiceus*	2											
Passeriformes	Tyrannidae	*Elaenia*	*flavogaster*							1					
Passeriformes	Tyrannidae	*Mionectes*	*oleagineus*		1	1				1				1	1
Passeriformes	Tyrannidae	*Myiarchus*	*tuberculifer*			1									
Passeriformes	Tyrannidae	*Myiozetetes*	*granadensis*							1					
Passeriformes	Tyrannidae	*Myiozetetes*	*similis*			3									
Passeriformes	Tyrannidae	*Onychorhynchus*	*coronatus*							1					
Passeriformes	Tyrannidae	*Platyrinchus*	*coronatus*		1										
Passeriformes	Tyrannidae	*Tolmomyias*	*sulphurescens*							1					
Passeriformes	Vireonidae	*Hylophilus*	*flavipes*	1											
Piciformes	Galbulidae	*Galbula*	*ruficauda*							2					
Piciformes	Picidae	*Melanerpes*	*pucherani*							1					
Piciformes	Ramphastidae	*Pteroglossus*	*torquatus*						1						
Trogoniformes	Trogonidae	*Trogon*	*massena*									1			
Trogoniformes	Trogonidae	*Trogon*	*rufus*		1							2			

6. CYAFLOMIO: all individuals from the three species that were sampled from both Costa Rica and Peru: *Cyanoides cyanocompsa*, *Florisuga mellivora* and *Mionectes oleaginous*. This allows investigation of large-scale continental differences between birds of the same species. (N=17)
7. MANAKINS: all individuals from two species that were sampled multiple times and belong to the same family: *Manacus candei* and *Pipra mentalis*. This dataset allowed us to look at the differentiation between closely related species. (N=13)
8. TIRIMBINA: all individuals from Tirimbina, the most densely sampled locality. This removes geographic variation from the dataset. (N=45)

5.2.3. Taxonomic Assignment and Clustering Analyses

The microbial ecology package QIIME (Caporaso et al 2010) was used for the following analyses. First, the reads were assigned to phylotypes at 97% sequence similarity because 3% is frequently cited as the "species" level of microbial taxonomy (Schloss and Handelsman 2005), hereinafter "phylotypes". Next, we assigned taxonomies to phylotypes using the RDP CLASSIFIER PROGRAM (Wang et al 2007), with the default confidence threshold of 80%. A "core microbiota" was calculated and included all phylotypes that were found in 100% of the samples.

A pairwise UniFrac (Lozupone and Knight 2005) distance matrix (UDM) was constructed between each gut microbial community (i.e., each bird specimen). UniFrac distances are

calculated based on the amount of branch length in a phylogenetic tree that is unique to either of two environments (versus how much of the tree is shared by the environments). These distances can be based on presence-absence of OTUs ("unweighted") or weighted by abundance. Our microbial phylogenetic tree was constructed with FASTTREE (Price et al 2009). To reduce the effects of sampling (sequencing) bias, all individuals were randomly reduced to 3 652 reads, equal to the lowest number of reads for any bird in the dataset.

We constructed UPGMA dendrograms based on both the unweighted UDM and weighted UDM to visually represent the relatedness of the gut microbiota for all datasets. Principal coordinates analysis (PCoA) was also performed on both the weighted and unweighted UniFrac distance matrices.

As a complement to the phylogenetic-based methods, we visualized the data with nonmetric multidimensional scaling (NMDS). We square root-transformed the percentage of each sample that belonged to each bacterial phylum, then created a pairwise distance matrix using Bray-Curtis dissimilarity, applied through the VEGDIST function of the VEGAN package (Oksanen et al 2011) in R (R Development Core Team 2010). The NMDS function of the ECODIST package (Goslee and Urban 2007) was then used to calculate the two-dimensional positions of the samples (such that closer samples are more similar), the stress and R^2 value of the plot. Stress values >0.3 should not be considered valid whereas values <0.2 can be considered a good representation of the data (Quinn and Keough 2002).

5.2.4. Categorical Variable Significance

To look for a relationship between categorical variables associated with each bird and the microbial communities, we used the statistical tools ADONIS (McArdle and Anderson 2001) and ANOSIM (Clarke 1993) implemented in QIIME. The categorical variables included the current American Ornithologists' Union South American Classification Committee's taxonomy, i.e., order, family, genus and species (Remsen et al., Version 7 December 2012), ecological variables, including dietary specialization and habitat (Bennett & Owens, 2002), spatial variables and individual properties, like age (based on percent of skull ossification), stomach contents (e.g., "insects" or "plant material") and bacterial richness. Table 5.2 gives a detailed list of the variables and their sources. We calculated significance of all variables for both the weighted and unweighted UniFrac distance matrices with 999 iterations.

After testing the significance of each variable independently, we ran an additional Adonis test on the most frequently significant variables to quantify the amount of variation each variable was responsible for. We used the full dataset's unweighted and weighted UDMs as input, calculated 999 iterations, and permuted the order of the variables, which can affect the results of the test. Finally, we constrained the analyses to only permute the data within bird orders, as a measure of controlling for taxonomy. We then reran the weighted and unweighted UDMs.

Table 5.2. Categorical variables tested for significance, including the number of categories within each variable (Cat) and a list of the possible (except taxonomic categories).

Variable	Cat	Description and Source
Order	8	Bird order (Remsen et al. 2012)
Family	24	Bird family (Remsen et al. 2012)
Genus	53	Bird genus (Remsen et al. 2012)
Species	59	Bird species (Remsen et al. 2012)
Diet Specific	11	Specific dietary specialization: nectar, generalist, insect, seed, arthropod, fruit, insect/fruit, fruit/insect, nectar/insect, arthropod/vertebrates, fruit/nectar/insects (C. Sanchez, pers. com.)
Diet Broad	3	Broad dietary specialization: plant material, animal material, both
Diet B&O	4	Broad dietary specialization assigned by Bennett and Owens (2002): frugivore, nectarivore, insectivore, omnivore
Habitat	5	General habitat assigned by Bennett and Owens (2002): woodland, forest, forest/grassland, forest/grassland/scrub, all habitats
Foraging Strata	9	Foraging strata assigned by Stotz et al. (1996): canopy, midstory, understory, terrestrial, under/midstory, midstory/canopy, terrestrial/understory, terrestrial/midcanopy, understory/canopy
Locality	12	Sampling locality (see igure 5.1): Tirimbina, Londres, Los Charcos, Piedras Blancas, Golfo Dulce, El Copal, Janirvan, Tuba Creek, San Jorge II, La Selva, Veragua, Santa Juana
Country	2	Country of sampling: Costa Rica, Peru
NSEW	6	Relative localition of sampling locality: north-east Costa Rica, midwest Costa Rica, southwest Costa Rica, middle Costa Rica, mideast Costa Rica, Peru
Elevation	13	Elevation of sampling locality: 65m, 75m, 80m, 110m, 170m, 200m, 250m, 260m, 325m, 400m, 415m, 430m, 1050m
Sex	3	Sex of the bird: male, female, unknown
Age	14	Percent of skull ossification (a proxy for age of bird): 0, 3, 5, 10, 15, 20, 25, 50, 70, 75, 90, 95, 100, unknown
Stomach Contents	12	Contents of stomach at time of collection: insects, seeds, fruit, plant material, seeds/insects, insects/pollen, fruit/insects, insects/plants, seeds/plants, fruit/insects/seeds, empty, unknown
Phyla Richness	10	Number of bacterial phyla identified in the gut microbiota fingerprint: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17
Species Richness	5	Which fifth the number of 97%OTUs identified in a gut microbiota fingerprint belongs to: 20%ile, 40%ile, 60%ile, 80%ile, 100%ile

5.2.5. Spatial and taxonomic tests

We compared the weighted and unweighted UDM to a pairwise matrix of geographic distance between each of the samples using a Mantel test, with 999 permutations, to look for a concurrent increase in microbial distance as geographic distance increased (isolation by distance). We also tested whether the occurrence of the bacterial phyla differed in the Costa Rica and Peru samples, using a student's T-test and a significance level of 0.05. We tested both "weighted" occurrence data (the average percent composition of each phyla across all individuals) and "unweighted" occurrence (the percent of individuals that contained at least one representative from the phylum). We also constructed a heatmap of sampling localities vs. bacterial phyla; these data were the proportion of individuals at each locality that contained each phylum. To visually

inspect the distribution of bacterial phylotypes across host taxonomy and sampling locality, we constructed heatmaps where the cells were colored by relative abundance.

5.3. RESULTS

After removing some sequence reads during initial quality control steps, 9 897 718 pairs of reads remained with no errors in priming sequence, region of overlap or individual tags. Eleven hundred and sixty seven potentially chimeric sequences (0.01% of reads) were then discarded. A further 358 725 sequences that did not align to the domain Bacteria (3.6% of reads) were removed; 75% of these discarded reads belonged to 11 individuals. The reads passing these filters were included in the final dataset, totaling 9 537 817 sequences and averaging 82 222 sequences per individual; reads/sample varied by over two orders of magnitude (range: 3 652 – 853 078).

Four bacterial phyla were detected in all individuals: Proteobacteria, Firmicutes, Bacteroidetes and Actinobacteria comprising an average of 46.3%, 37.3%, 3.3% and 1.4% of each sample, respectively (Fig. 5.2). An additional 16 phyla were identified: Acidobacteria, Chlamydiae, Chloroflexi, Cyanobacteria, Deinococcus-Thermus, Fusobacteria, Lentisphaerae, Nitrospira, OD1, OP10, OP11, Planctomycetes, Spirochaetes, TM7, Tenericutes, Verrucomicrobia. An average of 10.6% of sequences from each individual were from unknown phyla within Bacteria. The core microbiota contained 56 phylotypes, 32 of which aligned to 26 known genera (Table 5.3). An additional 48 phylotypes were detected in >95% of the samples. The number of species-level phylotypes per bird varied between 109 and 288, with an average of 201 (standard deviation = 35). Replicate samples were similar to one another in taxonomic composition (Fig. 5.2) and generally clustered close to one another in multivariate space (data not shown).

5.3.1. Clustering Analyses

The PCoA for the unweighted UniFrac distance matrix displayed clustering by taxonomic order (Fig. 5.3), but little obvious clustering by foraging stratum, diet or sampling locality. Host order displayed the most clustering in the weighted UniFrac distance matrix PCoA and NMDS plot as well (Fig. 5.3).

5.3.2. Categorical Variable Significance

To look for correlations between categorical variables and the gut microbiota, we conducted two statistical tests for significance on both the weighted and unweighted UDMs. We then constructed histograms of how frequently these four tests were significant at $p<0.05$ across the eight datasets (Fig. 5.4A), then summarized the results across all datasets (Fig 5.4B). Taxonomic variables were the most frequently significant, with all four categories (order, family, genus, species) being significant in over 50% of the tests. The three dietary variables were also frequently significant; the broad dietary specialization category ("Diet Broad") and the Bennett and Owens diet variable ("Diet B&O") were both more frequently significant than host genus and species across datasets (Fig. 5.4B). The other ecological variables, foraging stratum and habitat, were significant in 43% and 46% of the tests, respectively. The various locality variables were all significant in less than 18% of the tests. Sex, age and stomach contents were significant

Table 5. 3. Bacterial taxa identified in 100% of the bird samples (assigned by RDP Classifier Program). Gray boxes indicate the phylotype did not align to any named taxa at that taxonomic rank. Numbers in boxes are the total number of identified phylotypes in that group. Full length gray bar at bottom indicates a phylotype that did not align to any named phylum.

Phylum	Class	Order	Family	Genus
Actinobacteria 5	Actinobacteria 5	Actinomycetales 4	Corynebacteriaceae 1	*Corynebacterium* 1
			Microbacteriaceae 1	
			Propionibacteriaceae 1	*Propionibacterium* 1
Bacteroidetes 5	Bacteroidia 1	Bacteroidales 1	Bacteroidaceae 1	*Bacteroides* 1
	Flavobacteria 3	Flavobacteriales 3	Flavobacteriaceae 3	*Chryseobacterium* 1
				Planobacterium 1
Firmicutes 17	Bacilli 12	Bacillales 4	Bacillaceae 2	*Anoxybacillus* 1
			Staphylococcaceae 1	*Staphylococcus* 1
		Lactobacillales 7	Lactobacillaceae 1	*Lactobacillus* 1
			Leuconostocaceae 2	*Leuconostoc* 1
				Weissella 1
			Streptococcaceae 3	*Lactococcus* 1
				Streptococcus 1
	Clostridia 4	Clostridiales 4	Clostridiaceae 1	*Clostridium* 1
			Veillonellaceae 2	*Veillonella* 1
Proteobacteria 27	Alpha-proteobacteria 4	Rhizobiales 2	Methylobacteriaceae 1	*Methylobacterium* 1
		Sphingomonadales 1	Sphingomonadaceae 1	*Sphingomonas* 1
	Beta-proteobacteria 9	Burkholderiales 8	Burkholderiales incertae sedis 2	*Aquabacterium* 1
				Tepidimonas 1
			Comamonadaceae 4	*Acidovorax* 1
				Diaphorobacter 1
				Schlegelella 1
			Oxalobacteraceae 1	*Janthinobacterium* 1
	Epsilon-proteobacteria 2	Campylobacterales 2	Campylobacteraceae 1	*Campylobacter* 1
			Helicobacteraceae 1	*Helicobacter* 1
	Gamma-proteobacteria 11	Aeromonadales 1	Aeromonadaceae 1	*Aeromonas* 1
		Enterobacteriales 4	Enterobacteriaceae 4	*Escherichia* 1
				Kluyvera 1
				Yersinia 1
		Pseudomonadales 3	Moraxellaceae 2	*Acinetobacter* 1
				Enhydrobacter 1
			Pseudomonadaceae 1	*Pseudomonas* 1
		Xanthomonadales 2	Xanthomonadaceae 2	*Stenotrophomonas* 1

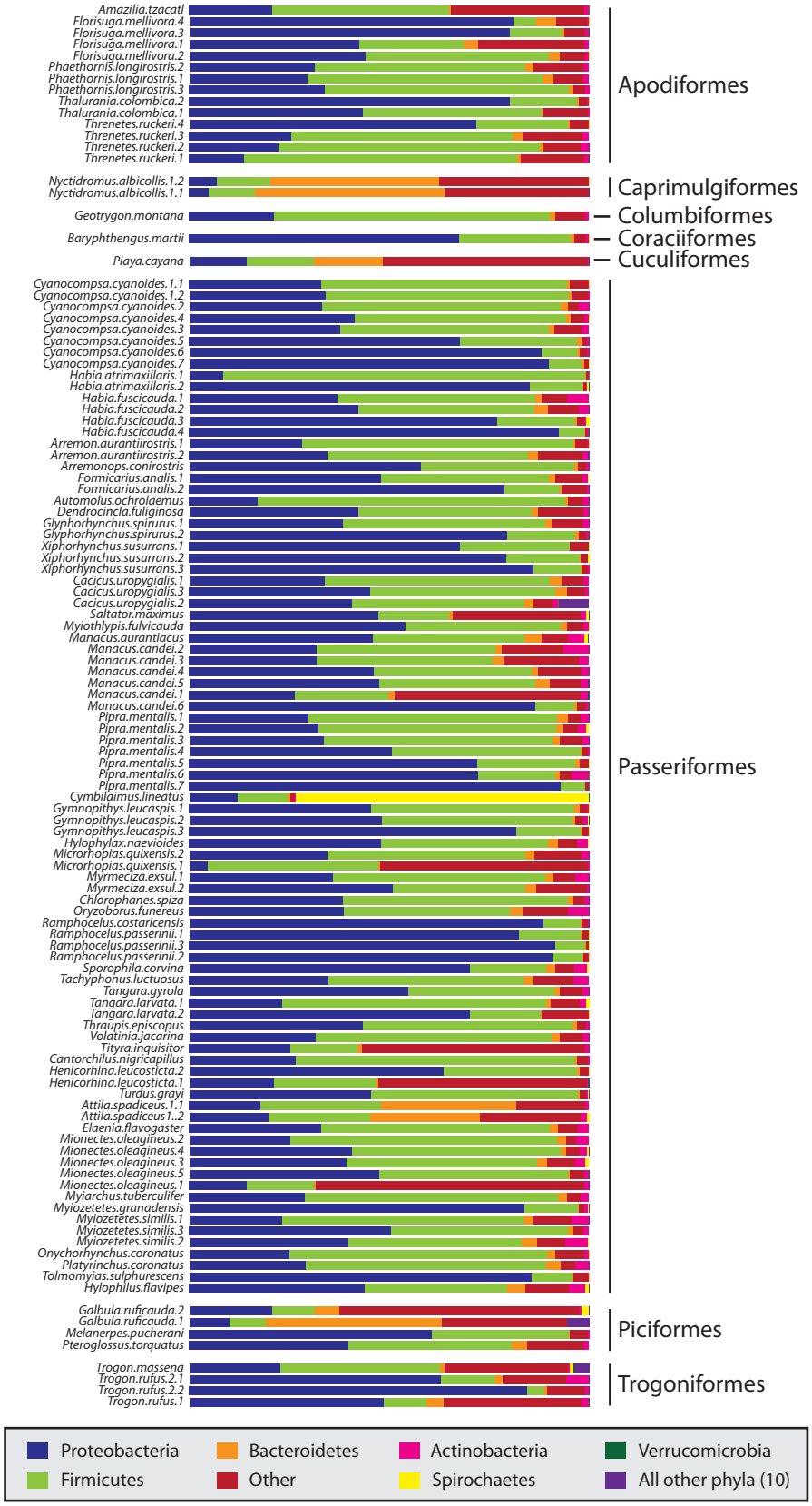


Figure 5.2. Relative contribution of seven bacterial phyla to each of the samples, separated by bird order.

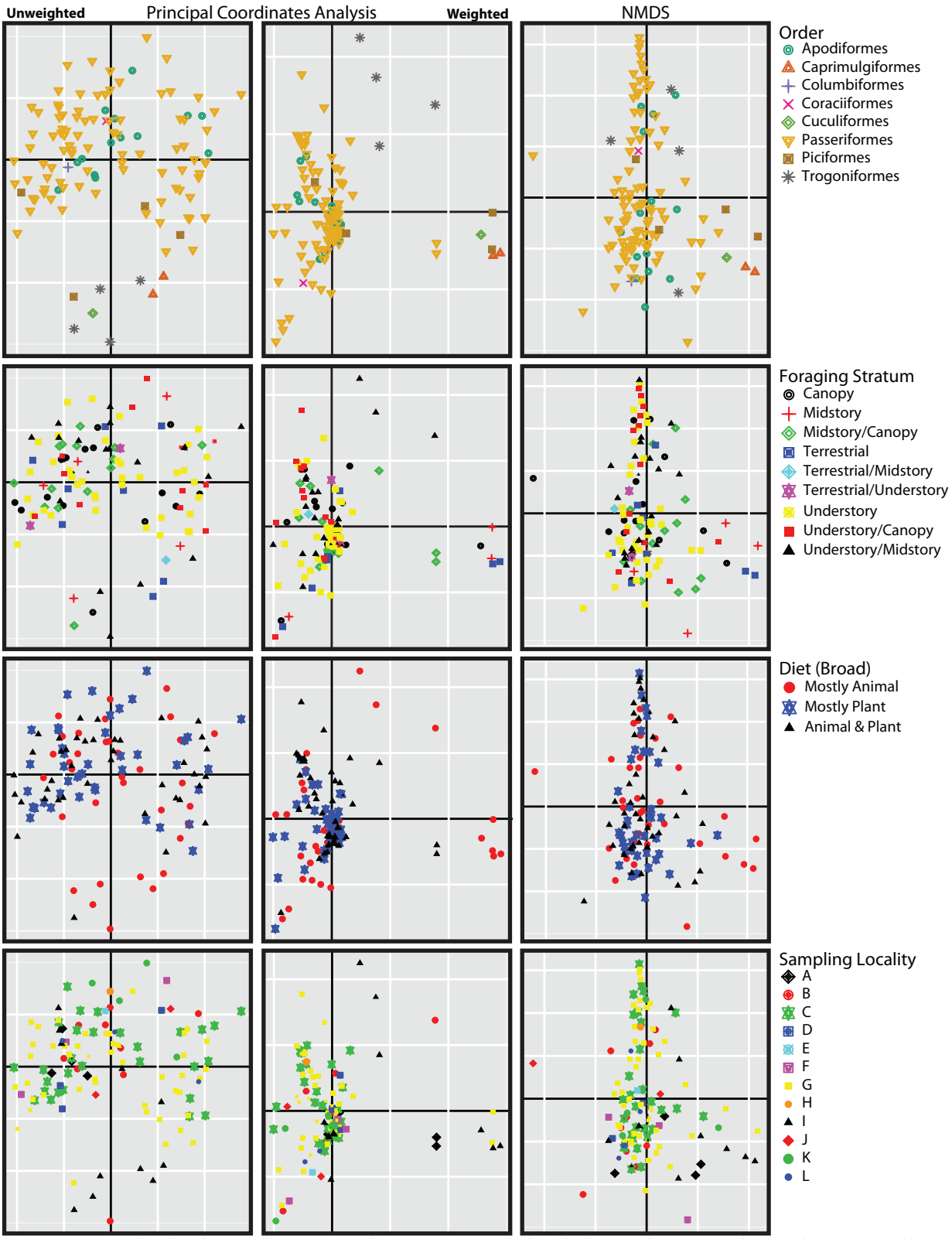


Figure 5.3. Principal coordinates analyses on unweighted (left column) and weighted (middle column) UniFrac distances and NMDS analysis of bacterial composition of samples (right column). Samples are colored by bird order (top row), foraging strata (second row), diet (third row) and sampling locality (bottom row).

in less than 10% of the tests. Bacterial richness was significant in 28% and 21% of the tests, at the phylum and species levels, respectively.

In the multifactor Adonis test, "Foraging Stratum" explained the most amount of variation for both the unweighted and weighted UniFrac distance matrices (Table 5.4), although the *p*-values were rather different (unweighted UDM p=0.004 and weighted UDM p=0.116). Bacterial phyla richness (p=0.016) and host order (p=0.002) were also significant in the unweighted analysis, accounting for 2% and 10% of the variation, respectively. No variables were significant (p<0.05) in the weighted analysis. When data were permuted within the taxonomic orders (i.e., controlling for taxonomy), the significance of the variables did not change (Table 5.4C and 5.4D).

5.3.3. Spatial and taxonomic tests

There was no signal for isolation by distance in the FULL DATASET unweighted (R statistic = 0.04405, p = 0.525) or weighted UDM (R statistic = -0.06219, p = 0.462). A scatterplot of geographic distance vs. UniFrac distance (Fig. 5.5) showed no correlation between the distances. The relationship was also not significant within the single species dataset, *CYANOIDES*, (unweighted UDM: R statistic = 0.09876, p = 0.572; weighted UDM: R statistic = 0.20815, p = 0.212).

We also tested for differences between Costa Rica and Peru in the occurrence of bacterial phyla. The "weighted" occurrence data (including abundance information) were not significantly different between the countries (Fig. 5.6A). The "unweighted" occurrence data (presence/absence data) indicated a significant difference between three bacterial phyla (Fig. 5.6B): Chloroflexi (p=0.033), Cyanobacteria (p=0.036), OP11 (p=0.008). The percentage of individuals with unclassified sequences was also significant (p=8×10^{-7}). The heatmap of presence of bacterial phyla across the sampling localities revealed no locality specific phyla (Fig. 5.6C).

The heatmap of bacterial phylotypes vs. host taxonomy revealed little clustering and showed how specific phylotypes were found in high abundance in most individuals (Fig. 5.7); most these phylotypes belonged to the Firmicutes and Proteobacteria. When the data were rearranged in order of sampling locality, there was no visible clustering (data not shown).

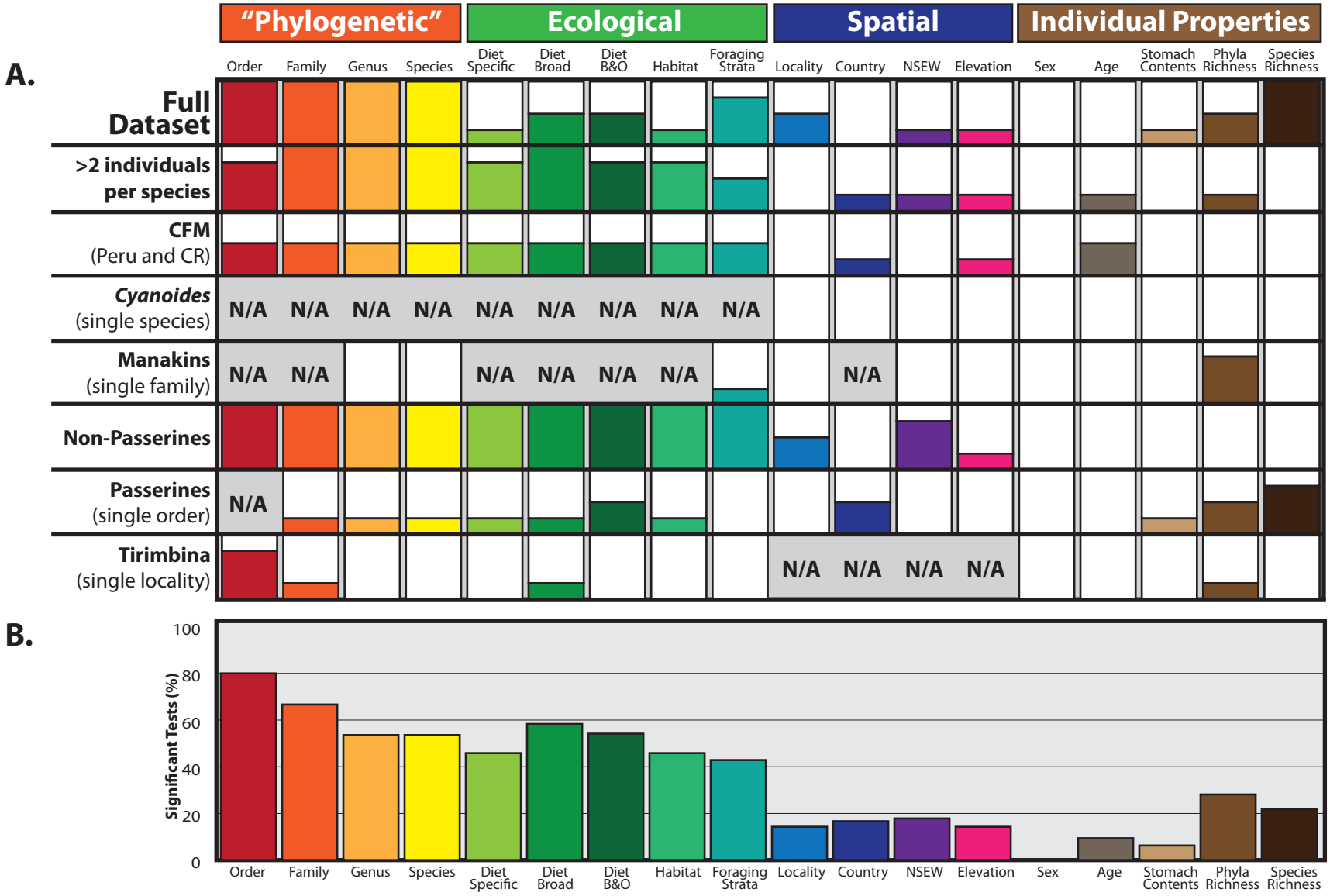


Figure 5.4. Histograms of how frequently each categorical variable was significant for each dataset (A) and percentage of significant tests across applicable datasets (B). Details on variables given in Table 5.2; datasets described in 5.2. Methods.

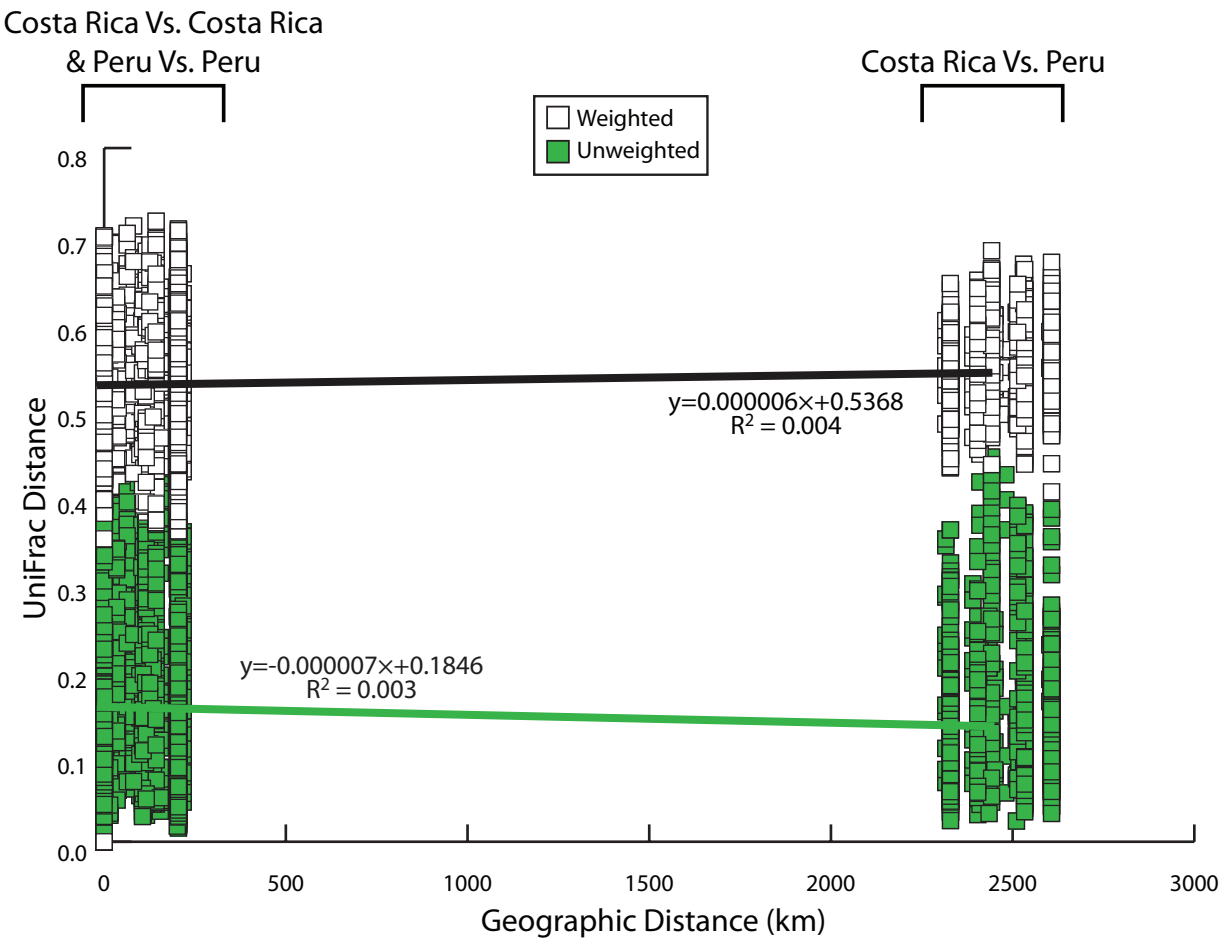


Figure 5.5. Scatterplot of pairwise UniFrac distances against pairwise geographic distances between all samples. Unweighted UniFrac distances in green, weighted UniFrac distances in white. Trendlines with their equations and associated R^2 values also shown.

Table 5.4. Multifactorial Adonis tests.

	Df	Sum Of Squares	Mean Squares	F.Model	R^2	Pr(>F)
A. Unweighted UniFrac Distance Matrix						
Diet	2	0.002	0.001	0.747	0.012	0.660
Foraging Stratum	8	0.022	0.003	1.787	**0.111**	**0.004**
Bacterial Phyla	1	0.005	0.005	3.072	**0.024**	**0.016**
Order	7	0.020	0.003	1.833	**0.100**	**0.002**
Residuals	97	0.149	0.002		0.754	
Total	115	0.197			1.000	
B. Weighted UniFrac Distance Matrix						
Diet	2	0.022	0.011	0.418	0.007	0.805
Foraging Stratum	8	0.314	0.039	1.518	0.100	0.116
Bacterial Phyla	1	0.010	0.010	0.370	0.003	0.688
Order	7	0.292	0.042	1.614	0.093	0.080
Residuals	97	2.505	0.026		0.797	
Total	115	3.141			1.000	
C. Unweighted UniFrac Distance Matrix - Controlling for host order						
Diet	2	0.002	0.001	0.707	0.012	0.894
Foraging Stratum	8	0.022	0.003	1.693	**0.111**	**0.009**
Bacterial Phyla	1	0.005	0.005	2.909	**0.024**	**0.024**
Residuals	104	0.168	0.002		0.853	
Total	115	0.197			1.000	
D. Weighted UniFrac Distance Matrix - Controlling for host order						
Diet	2	0.022	0.011	0.402	0.007	0.790
Foraging Stratum	8	0.314	0.039	1.458	0.100	0.303
Bacterial Phyla	1	0.010	0.010	0.355	0.003	0.702
Residuals	104	2.797	0.027		0.890	
Total	115	3.141			1.000	

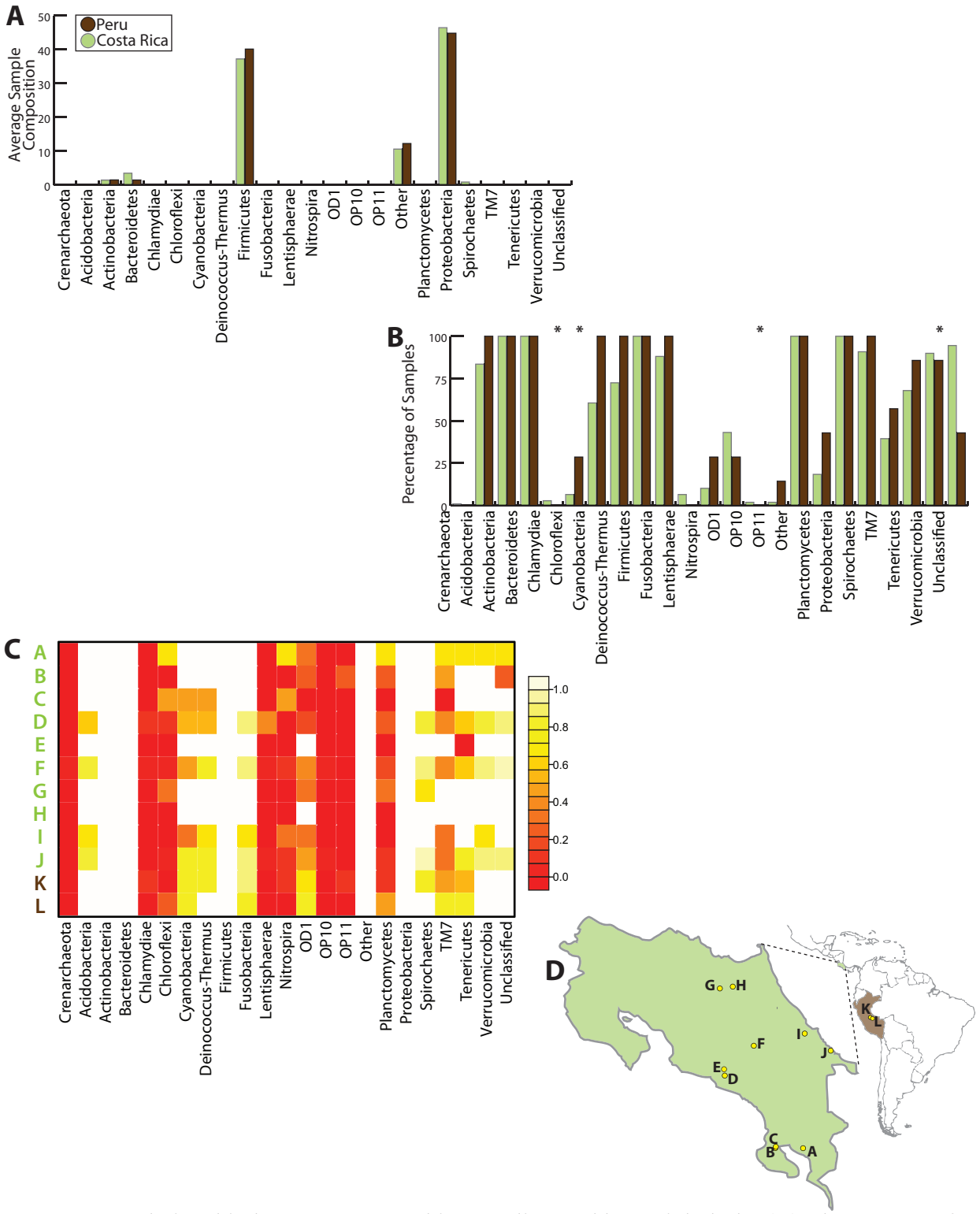


Figure 5.6. Relationship between geographic sampling and bacterial phyla. (A) The percent of each sample that is comprised by each bacterial phyla, averaged by country. (B) The percentage of samples that contained each phylum for each country. (C) Heatmap of how frequently each phylum was found in the birds from each sampling locality (D). Asterisks denote a significant difference between the Costa Rica (green) and Peru (brown) samples.

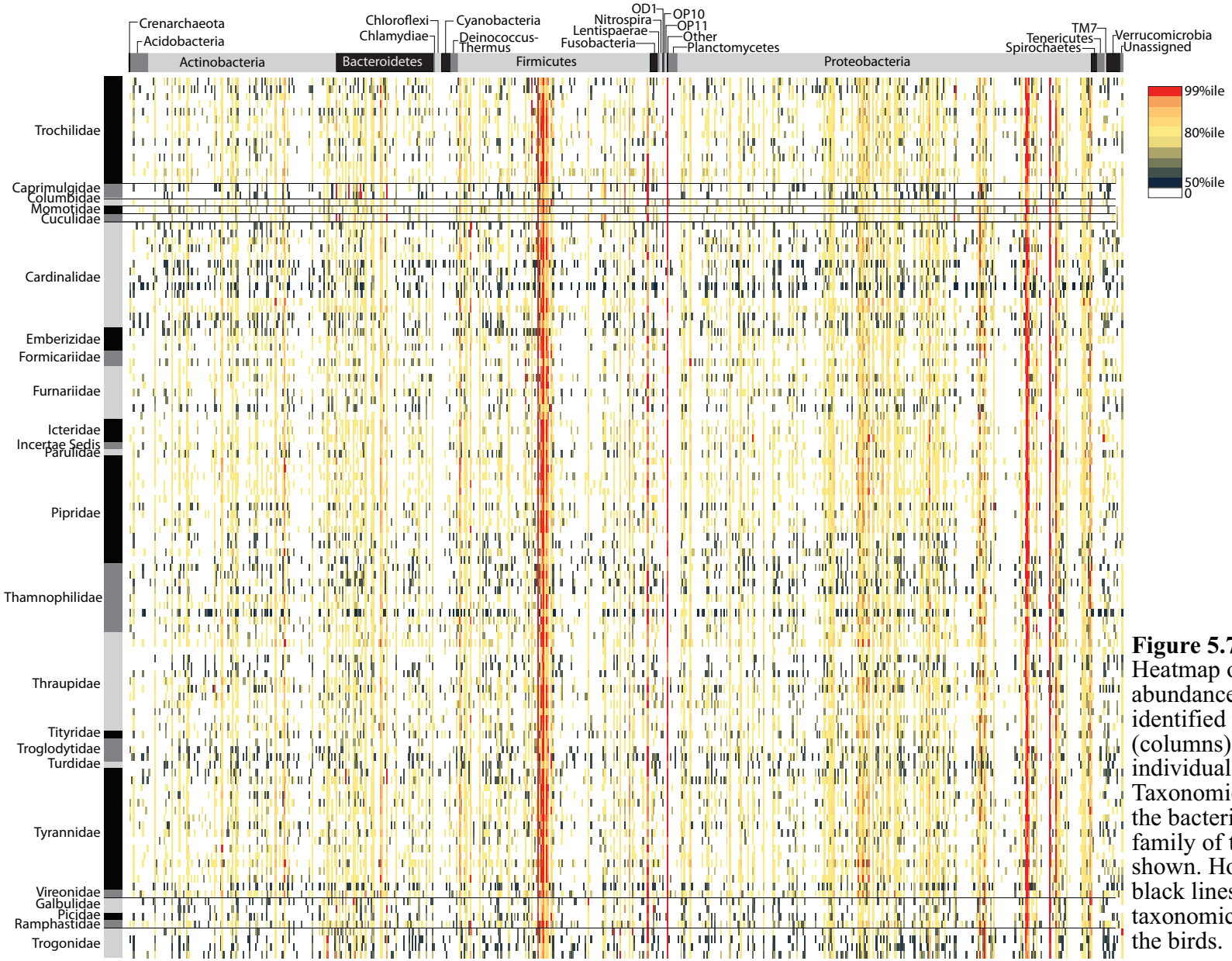


Figure 5.7. Heatmap of relative abundance of each identified phylotype (columns) for each individual (rows). Taxonomic class of the bacteria and family of the birds is shown. Horizontal black lines delimit taxonomic orders of the birds.

5.4. DISCUSSION

The primary aim of this study was to assess the utility of gut microbiota fingerprints for phylogeographic inference. Phylogeography combines genetics and biogeography to investigate the spatial distribution of genetic lineages; therefore, to be suitable for this use, the gut microbiota must contain phylogenetic and/or temporal or spatial information.

There is clear evidence for associations between gut microbiota and host taxonomy. It is unknown whether it is actually genetic distance that is important and further experimentation linking genetic divergence on an individual scale with microbiota divergence would be most interesting. All the locality variables had poor correlations with the gut microbiota and our spatial tests revealed no statistical significance between space and microbiota. Can we therefore conclude that physical space has no effect on the microbiota? Perhaps. The whole communities appear to not be more closely related the closer they are together, both across Aves and within *Cyanocompsa cyanoides*. Alternatively, the apparent lack of effect of locality may be an issue of sampling. This dataset contains few species (or even genera) with multiple individuals, i.e., much less than phylogeographic level sampling. If locality is important, we might expect it to be working within species instead of across higher taxonomic levels. In Chapter 4, brown-headed cowbirds displayed a geographic effect, but that study contained 34 samples from the species. Scale of analysis may be critical for detecting what factors are contributing to divergence.

We also found associations between the gut microbiota and host ecology. Dietary specialization and not the contents of the stomach were significantly associated with the microbiota – broad dietary classifications (mostly plant, mostly animal, plant/animal) were more significant than specific dietary specializations. This implies long term habits or nutritional content have greater influence than day to day food intake, which is consistent with other studies showing the stability of the avian gut community once established (Benskin et al., 2010). Of course, the "stomach contents" variable contains a lot of variance particularly with respect to specificity and accuracy, as it is recorded in the field and only general data are taken (i.e., "plant material" or "insects"). Additionally, many stomachs are empty since birds caught in mist nets frequently evacuate their bowels. Although a phylogenetic effect between taxonomy and broad dietary specialization is possible, each bird order with greater than two samples contained multiple specializations.

Strata and habitat are important aspects of avian biology. Foraging strata is associated with genetic divergence in Neotropical birds (Burney & Brumfield, 2009) because ecology affects dispersal ability. These results reinforce the importance of foraging strata and support the important role that ecology can play in differentiation of both host and microbiota. Microbes from the same ecological niche on the human body are able to share genes on a global scale (Smillie et al., 2011) – perhaps the microbiota of birds that share ecological niches are able to transfer genetic material as well. The importance of external environment may be higher in birds than in other taxa. Mammals are inoculated with complex microbial communities at birth by delivery through the vaginal canal (Mandar & Mikelsaar, 1996, Palmer et al., 2007). Birds, on the other hand, hatch from eggs and their initial exposure is to the environment, particularly nests and eggshells (Kohl, 2012). The strength of vertical association between generations of birds remains to be assessed and would be an important comparison to other vertebrate microbiota.

We have learned more about microbes than their avian hosts with these data. Within the genera found in all birds were *Lactococcus*, *Streptococcus* and *Campylobacter* bacteria that have been frequently sampled in domesticated poultry (Olsen et al., 2008, Qu et al., 2008, Scupham, 2009, Lan et al., 2005). The concept of a "core microbiota" has been cast in doubt recently since interindividual variation is so great and studies employing deep sequencing rarely recover identical phylotypes across all samples (Lozupone et al., 2012). However, higher taxonomic patterns are frequently consistent within a species (Eckburg et al., 2005). Bacterial richness at both the phylum and species levels is correlated with the gut data (Fig. 5.4), perhaps indicating that some aspect of the microbial communities is important for final gut microbiota structure. This study relies on a single marker (V6) and taxonomic consistency is extraordinarily rare at certain taxonomic scales, even in closely related individuals (Turnbaugh et al., 2010). Functional consistency, on the other hand, is common and methods that evaluate the protein families may add great resolution to these results.

Although there is not strong evidence for a spatial component recapitulated in the gut microbiota, there are correlations with host taxonomy and ecology, which merit the continued investigation of microbiota as a phylogeographic marker. The sampling may be obscuring the effect that physical space is having on the microbiota; additional sampling may correct this. It is also worth noting that these data were obtained without additional sampling effort and these methodologies could be incorporated into field protocols, as all information gleaned represents data gained on the specimen. Phylogeography aims to understand the spatial arrangement of the biodiversity around us – expanding the discipline into the microbiota is relatively unexplored avenue for biological inference. Investigating symbiotic relationships across spatial, taxonomic and ecological scales is an exciting avenue for synergistic research across traditional disciplines and increased understanding of the natural world.

Chapter 6.
Conclusions

I am drawn to DNA sequence data for many reasons. All the life on this planet is comprised of the same four base pairs, making sequence data very simple in one respect. On the other hand, one organism's genome can be billions of bases long, and different loci therein can have different evolutionary histories, making understanding DNA data very complex. Biologists are united by the theory of evolution, which can explain much of the complexity we see. High-throughput sequencing (HTS) facilitates the gathering of genome-scale data with relative ease and minimal cost. Applying HTS to traditional evolutionary biology questions requires new bioinformatics tools. Furthermore, HTS allows us to explore novel biological questions, including the extremely complex and important consortia of microorganisms living in vertebrates. These are the two main avenues of my research and this dissertation: bioinformatics tools and gut microbiota of Neotropical birds.

In Chapter 2, I designed, wrote and tested code for a program (PRGMATIC) that transforms raw HTS data into phased diploid loci. With this program, researchers interested in phylogenetics, phylogeography and population biology can use HTS data with minimal investment into bioinformatics. Using simulations, I showed this program was capable of correctly calling several common types of genetic polymorphism (Table 2.2). Chapter 3 is a companion program to PRGMATIC, called LOCINGS, that quickly and easily displays the most pertinent parameters for phylogenetic inference: coverage and number of SNPs. Using LOCINGS, users can also export the raw data used to call a locus and with these data, a researcher can investigate whether a locus is single copy or multiple copy, an important and unresolved aspect of anonymous data. Together these programs have been used to infer population structure in pitcher plants (Zellmer et al., 2012), describe a hybrid zone in rails (Maley & Brumfield, 2013) and construct a species tree in birds (McCormack et al., 2012).

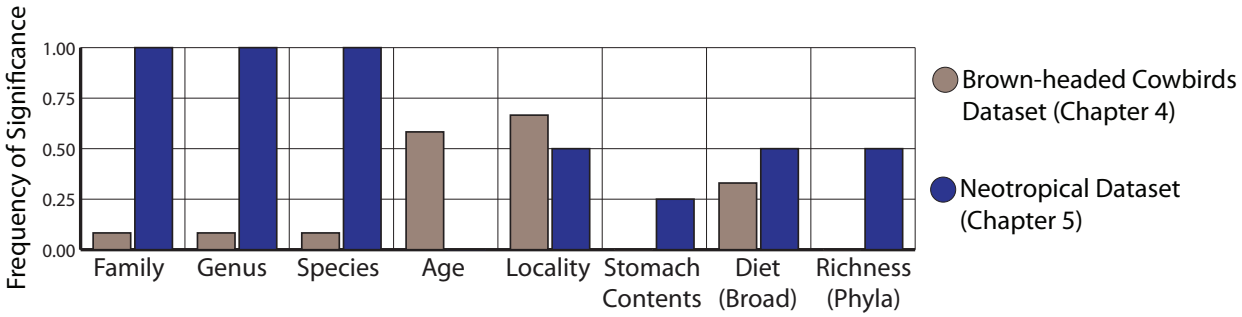


Figure 6.1. Comparison of categorical variable significance for the two bird datasets. Original data displayed in Figs. 4.6 and 5.4 and methods outlined in 4.2.4 and 5.2.4. All categories were consistent across the two datasets although number of tests differed.

The chapters on bird gut microbiota found complimentary results. With the cowbirds (Chapter 4), space and age of the bird had the highest correlation with gut microbiota (Fig. 6.1). With the Neotropical birds, it was taxonomy and ecology of the host that were most correlated. While these results seem contradictory, it may be the scale of the sampling that determines the detectable signal. Whereas greater than 30 cowbird samples were used, no single species reached more than eight samples in the Neotropical dataset, which had substantially higher taxonomic

and ecological diversity. Chapter 4 was designed to identify signal within a species; Chapter 5 was designed to compare across species. Together, these two findings imply that gut microbiota may in fact be an appropriate marker for phylogeographic inference, since they contain both spatial and phylogenetic information. Further work is needed to explore optimal sampling strategies for a given set of hypotheses.

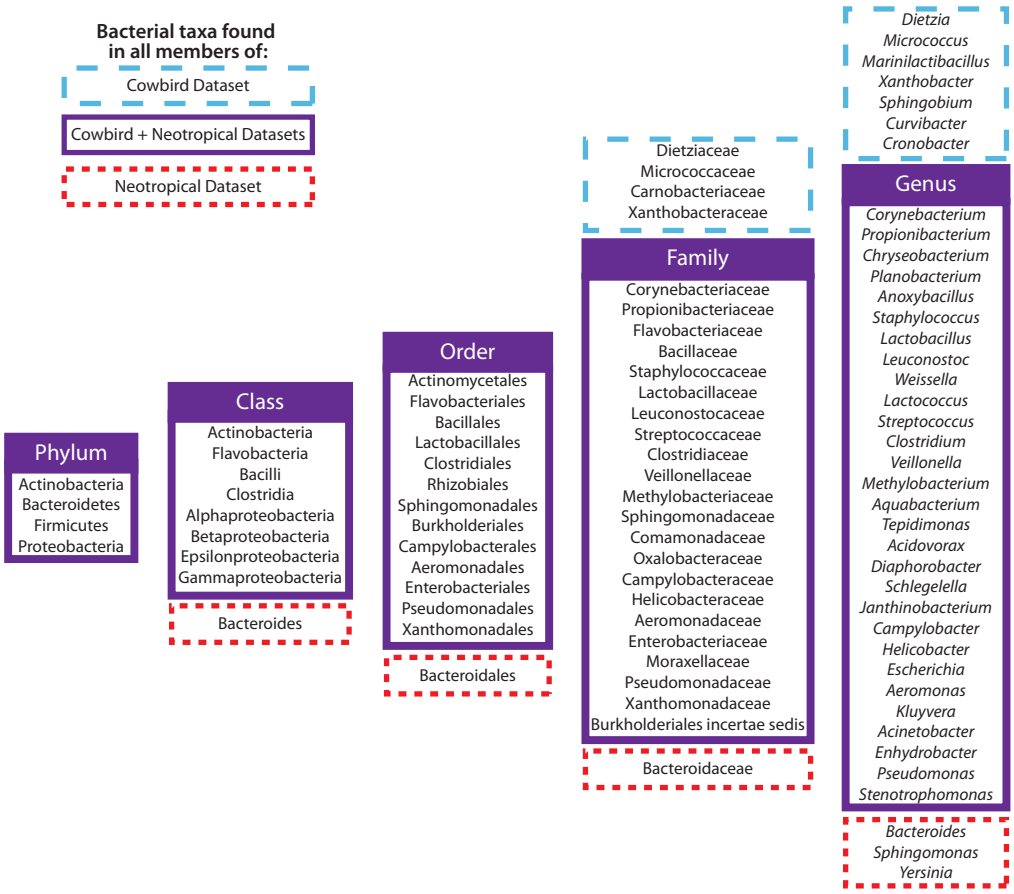


Figure 6.2. Comparison of core microbiome across all birds. Bacterial taxa shared by all birds in just the cowbird dataset (Chapter 4) outlined by large dash (blue); those shared by all birds in just the Neotropical birds dataset (Chapter 5) are outlined by small dashed line (red); those found in all birds inboth datasets are outlined by a solid line (purple).

This is also the first catalogue of Neotropical bird gut microbiota. I have shown that bird guts contain mostly Proteobacteria and Firmicutes (Figs. 4.2 and 5.2) and that the core microbiome across the datasets contains many taxa (Fig. 6.2), even at the genus level. These shared taxa include many genera broadly associated with gut microbiota across taxa and specifically found in birds, including *Campylobacter, Escherichia, Clostridium* and *Streptococcus* (Olsen et al., 2008, Qu et al., 2008, Scupham, 2009, Lan et al., 2005). One bacterial class was identified in all the individuals of the Neotropical bird dataset (Bacteroides); it would be interesting to investigate the cause of this difference, which could include major biological differences between the birds in the two datasets or sequencing error, among others.

I have also put bird gut microbiota variation in a broader biological context by comparing all the avian samples from this dissertation with the mammals and insect samples from Chapter 4

(Appendices D, E, F). Using the same methods outlined in Chapter 4.2.5, it is apparent that bird guts are distinct from other organisms and that they contain much variation (Fig. 6.3). Some of the insects and mammals cluster closer to the birds than they do to the rest of their respective classes. Further investigation into what traits these individuals share would be elucidating.

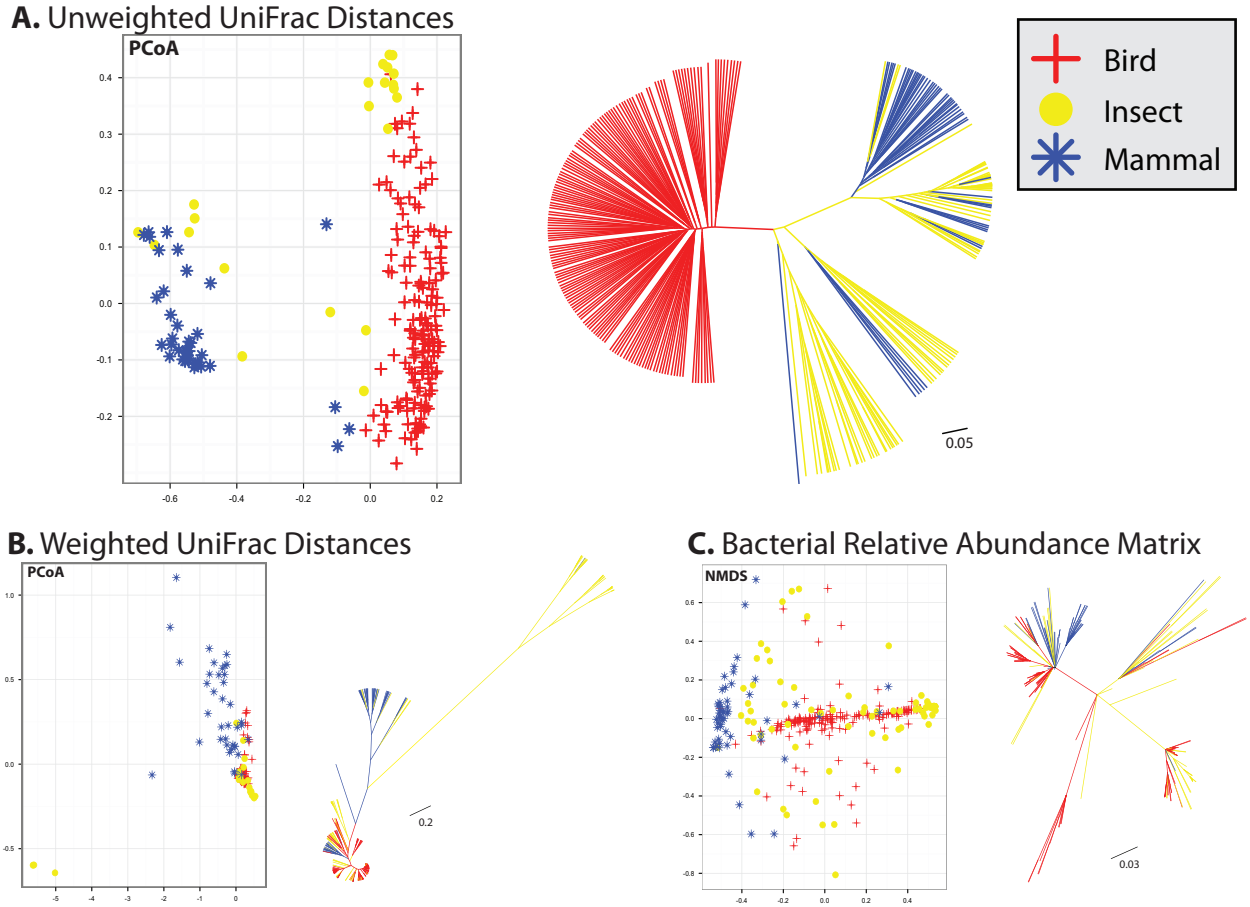


Figure 6.3. Visualizations of gut microbiota relatedness between birds, insects and mammals, including two dimensions of principal coordinates analyses on unweighted (A) and weighted (B) UniFrac distance matrices, as well as NMDS on the bird-by-bacterial phyla relative abundance matrix (C). Trees are dendrograms of the same underlying data matrix.

My goal for this dissertation was to learn about bioinformatics through computational tool development while maintaining an empirical aspect to my research. I believe I have achieved this and will continue to pursue bioinformatics and the spatial arrangement of microorganisms using HTS. In the future I would like to incorporate more robust microbial methods (e.g., metagenomics) that allow me to compare not only taxonomic diversity of environmental samples but functional differences as well. (Stevens & Hume, 1995, Qin et al., 2010)(Stevens & Hume, 1995, Qin et al., 2010)

References

Absalan, F. & Ronaghi, M. 2007. Molecular inversion probe assay. *Methods in Molecular Biology* **396**: 16.

Altschul, S., Gish, W., Miller, W., Myers, E. W. & Lipman, D. 1990. Basic local alignment search tool (BLAST). *Journal of Molecular Biology* **215**: 403-410.

Baird, N., Etter, P., Atwood, T., Currey, M., Shiver, A., Lewis, Z., Selker, E., Cresko, W. & Johnson, E. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**: 3376.

Balzer, S., Malde, K., Lanzen, A., Sharma, A. & Jonassen, I. 2010. Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics* **26**: 6.

Beerli, P. & Felsenstein, J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**: 763-773.

Bennett, P. M. & Owens, I. P. F. 2002. *Evolutionary Ecology of Birds: Life Histories, Mating Systems and Extinction*. Oxford University Press, New York.

Benskin, C. M. H., Rhodes, G., Pickup, R. W., Wilson, K. & Hartley, I. R. 2010. Diversity and temporal stability of bacterial communities in a model passerine bird, the zebra finch. *Molecular Ecology* **19**: 5531-5544.

Benson, A. K., Kelly, S. A., Legge, R., Ma, F., Low, S. J., Kim, J., Zhang, M., Oh, P. L., Nehrenberg, D., Hua, K., Kachman, S. D., Moriyama, E. N., Walter, J., Peterson, D. A. & Pomp, D. 2010. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 18933-18939.

Bevins, C. L. & Salzman, N. H. 2011. The potter's wheel: The host's role in sculpting its microbiota. *Cellular and Molecular Life Sciences* **68**: 3675-85.

Binladen, J., Gilber, M., Bolback, J., Panitz, F., Bendixen, C., Nielsen, R. & Willerslev, E. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One* **2**: e197.

Blanco, G., Lemus, J. A. & Grande, J. 2006. Faecal bacteria associated with different diets of wintering red kites: influence of livestock carcass dumps in microflora alteration and pathogen acquisition. *Journal of Applied Ecology* **43**: 990-998.

Brito, P. H. & Edwards, S. V. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica* **135**: 439-455.

Brumfield, R. T., Liu, L., Lum, D. E. & Edwards, S. V. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, Manacus) from multilocus sequence data. *Systematic Biology* **57**: 719-731.

Burney, C. W. & Brumfield, R. T. 2009. Ecology predicts levels of genetic differentiation in Neotropical birds. *American Naturalist* **174**: 358-368.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., Mcdonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335-6.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J. H. 2011. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**: 171-182.

Chevreux, B., Wetter, T. & Suhai, S. (1999) Genome sequence assembly using trace signals and additional sequence information. In: *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*, Vol. 99. pp. 45-56.

Clarke, K. R. 1993. Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology* **18**: 117-143.

Clarridge III, J. E. 2004. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clinical Microbiology Reviews* **17**: 840.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M. & Tiedje, J. M. 2009. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research* **37**: D141-145.

Colman, D. R., Toolson, E. C. & Takacs-Vesbach, C. D. 2012. Do diet and taxonomy influence insect gut bacterial communities? *Molecular Ecology* **21**: 5124-5137.

Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I. & Knight, R. 2009. Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694-1697.

Degnan, P., Pusey, A., Lonsdorf, E., Goodall, J., Wroblewski, E., Wilson, M., Rudicell, R., Hahn, B. H. & Ochman, H. 2012. Factors associated with the diversification of the gut microbial communities within chimpanzees from Gombe National Park. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 13034-13039.

Denou, E., Rezzonico, E., Panoff, J., Arigoni, F. & Brussow, H. 2009. A mesocosm of *Lactobacillus johnsonii*, *Bifidobacterium longum*, and *Escherichia coli* in the mouse gut. *DNA and Cell Biology* **28**: 413-422.

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D. & Daly, M. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**: 491-498.

Dominguez-Bello, M. G. & Blaser, M. 2011. The human microbiota as a marker for migrations of individuals and populations. *Annual Review of Anthropology* **40**: 451-474.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T. & Wilson, A. (2012) Geneious v5.6.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T. & Wilson, A. (2010) Geneious v5.1.

Eckburg, P., Bik, E., Bernstein, C., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S., Nelson, K. & Relman, D. 2005. Diversity of the human intestinal microbial flora. *Science* **308**: 1635-1638.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 6.

Edmonson, M. N., Zhang, J., Yan, C., Finney, R. P., Meerzaman, D. M. & Buetow, K. H. 2011. Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics* **27**: 865-866.

Ezenwa, V. O., Gerardo, N. M., Inouye, D. W., Medina, M. & Xavier, J. B. 2012. Animal behavior and the microbiome. *Science* **338**: 198-199.

Frank, D. N. & Pace, N. R. 2008. Gastrointestinal microbiology enters the metagenomics era. *Current Opinion in Gastroenterology* **24**: 4-10.

Fraune, S. & Bosch, T. C. G. 2007. Long-term maintenance of species-specific bacterial microbiota in the basal metazoan Hydra. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 13146-13151.

Gao, F., Bailes, E., Robertson, D., Chen, Y., Rodenburg, C., Michael, S., Cummins, L., Arthur, L., Peeters, M. & Shaw, G. 1999. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**: 436-440.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**: 759-769.

Gloor, G. B., Hummelen, R., Macklaim, J. M., Dickson, R. J., Fernandes, A. D., MacPhee, R. & Reid, G. 2010. Microbiome profiling by Illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS One* **5**: 15.

Godoy-Vitorino, F., Goldfarb, K. C., Brodie, E. L., Garcia-Amado, M. A., Michelangeli, F. & Domínguez-Bello, M. G. 2010. Developmental microbial ecology of the crop of the folivorous hoatzin. *ISME J* **4**: 611-620.

Godoy-Vitorino, F., Leal, S., Diaz, W., Rosales, J., Goldfarb, K., Garcia-Amado, M., Michelangeli, F., Brodie, E. L. & Dominguez Bello, M. G. 2012. Differences in crop bacterial community structure between hoatzins from different geographical locations. *Research in Microbiology* **163**: 211-220.

Goecks, J., Nekrutenko, A., Taylor, J. & Team, T. G. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**.

Gompert, Z., Forister, M., Fordyce, J., Nice, C., Williamson, R. & Buerkle, A. 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. *Molecular Ecology* **19**: 2455-2473.

Gordon, D., Abajian, C. & Green, P. 1998. Consed: A graphical tool for sequence finishing. *Genome Research* **8**: 8.

Goslee, S. C. & Urban, D. L. 2007. The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software* **22**: 1-19.

Hafner, M. S. & Nadler, S. A. 1988. Phylogenetic trees support the coevolution of parasites and their hosts. *Nature* **332**: 258-259.

Hahn, D. C. & Smith, G. W. 2011. Life history trade-offs between longevity and immunity in the parasitic Brown-Headed Cowbird? *Open Evolution Journal* **5**: 8-13.

Heijtz, R. D., Wang, S., Anuar, F., Qian, Y., Bjorkholm, B., Samuelsson, A., Hibberd, M. L., Forssberg, H. & Pettersson, S. 2011. Normal gut microbiota modulates brain development and behavior. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 3047-3052.

Hercus, C. 2009. *www.novocraft.com*. Last accessed: January 2013.

Hey, J. 2010. Isolation with migration models for more than two populations. *Molecular Biology and Evolution* **27**: 905.

Hird, S., Brumfield, R. & Carstens, B. 2011a. PRGmatic: An efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional-reference genome'. *Molecular Ecology Resources* **11**: 743-748.

Hird, S. M. 2012. lociNGS: A lightweight alternative for assessing suitability of next-generation loci for evolutionary analysis. *PLoS One* **7**: e46847.

Hird, S. M., Brumfield, R. T. & Carstens, B. C. 2011b. PRGmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional-reference genome'. *Molecular Ecology Resources*.

Hird, S. M., Cardiff, S. W., Dittmann, D. L., Carstens, B. C. & Brumfield, R. T. In Review. Nature, nurture and the gut microbiota of the brood-parasitic Brown-headed Cowbird (*Molothrus ater*). *The ISME Journal*.

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A. & Cresko, W. A. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics* **6**: 23.

Huang, X. & Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Research* **9**: 10.

Huber, T., Faulkner, G. & Hugenholtz, P. 2004. Bellerophon: A program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**: 2317-2319.

Hugenholtz, P., Goebel, B. & Pace, N. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology* **180**: 4765.

Inoue, R. & Ushida, K. 2006. Development of the intestinal microbiota in rats and its possible interactions with the evolution of the luminal IgA in the intestine. *FEMS Microbiology Ecology* **45**: 147-153.

Józefiak, D., Rutkowski, A. & Martin, S. 2004. Carbohydrate fermentation in the avian ceca: A review. *Animal Feed Science and Technology* **113**: 1-15.

Kilner, R. M., Madden, J. R. & Hauber, M. E. 2004. Brood parasitic cowbird nestlings use host young to procure resources. *Science* **305**: 877-879.

Knowles, L. & Maddison, W. 2002. Statistical phylogeography. *Molecular Ecology* **11**: 2623-2635.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K. & Ding, L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**: 2283-2285.

Kohl, K. D. 2012. Diversity and function of the avian gut microbiota. *Journal of Comparative Physiology B* **182**: 591-602.

Koopman, M. M. & Carstens, B. C. 2011. The microbial phyllogeography of the carnivorous plant *Sarracenia alata*. *Microbial Ecology* **61**: 750-758.

Kumar, S., Carlsen, T., Mevik, B.-H., Enger, P., Blaalid, R., Shalchian-Tabrizi, K. & Kauserud, H. 2011. CLOTU: An online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. *BMC Bioinformatics* **12**: 182.

Lan, Y., Verstegen, M., Tamminga, S. & Williams, B. 2005. The role of the commensal gut microbial community in broiler chickens. *World's Poultry Science Journal* **61**: 95-104.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: R25.

Lee, S., Sung, J., Lee, J. & Ko, G. 2011. A comparison of the gut microbiota of healthy adult twins living Korea and United States. *Applied and Environmental Microbiology* **77**: 7433-7437.

Ley, R., Bäckhed, F., Turnbaugh, P., Lozupone, C., Knight, R. D. & Gordon, J. 2005. Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences* **102**: 11070-11075.

Ley, R., Hamady, M., Lozupone, C., Turnbaugh, P., Ramey, R., Bircher, J., Schlegel, M., Tucker, T., Schrenzel, M. & Knight, R. 2008a. Evolution of mammals and their gut microbes. *Science* **320**: 1647.

Ley, R., Lozupone, C., Hamady, M., Knight, R. & Gordon, J. 2008b. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* **6**: 13.

Ley, R., Peterson, D. & Gordon, J. 2006. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* **124**: 837-848.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 7.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Subgroup, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li, R., Li, Y., Kristiansen, K. & Wang, J. 2008. SOAP: Short oligonucleotide alignment program. *Bioinformatics* **24**: 713-714.

Lowther, P. E. 1993. *Brown-headed Cowbird (Molothrus ater)*. American Ornithologists Union, Ithaca.

Lozupone, C. & Knight, R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**: 8228.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. 2012. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**: 220-230.

Lunter, G. & Goodson, M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* **21**: 936-939.

Maddison, D. R., Swofford, D. L. & Maddison, W. P. 1997. NEXUS: An extensible file format for systematic information. *Systematic Biology* **46**: 590-621.

Maley, J. M. & Brumfield, R. T. 2013. Mitochondrial and next-generation sequencing data are used to infer phylogenetic relationships and species limits in the Clapper/King Rail (*Rallus longirostris* & *elegans*) complex. *Condor* **In press**.

Mamanova, L., Coffey, A., Scott, C., Kozarewa, I., Turner, E., Kumar, A., Howard, E., Shendure, J. & Turner, D. 2009. Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7**: 111-118.

Mandar, R. & Mikelsaar, M. 1996. Transmission of mother's microflora to the newborn at birth. *Biology of the Neonate* **69**: 30-35.

McArdle, B. H. & Anderson, M. J. 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* **82**: 290-297.

McCormack, J., Hird, S., Zellmer, A. J., Carstens, B. & Brumfield, R. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* **66**: 526-538.

McCormack, J., Maley, J. M., Hird, S., Derryberry, E., Graves, G. & Brumfield, R. T. 2012. Next-generation sequencing reveals phylogeographic structure and a species tree for recent bird divergences. *Molecular Phylogenetics and Evolution* **62**: 397-406.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**: 1297-1303.

McPherson, J. 2009. Next-generation gap. *Nature Methods* **6**: S2-S5.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. & Marshall, D. 2009. Tablet - next generation sequence assembly visualization. *Bioinformatics* **26**: 2.

Moodley, Y., Linz, B., Yamaoka, Y., Windsor, H., Breurec, S., Wu, J.-Y., Maady, A., Bernhöft, S., Thiberge, J.-M. & Phuanukoonnon, S. 2009. The peopling of the Pacific from a bacterial perspective. *Science* **323**: 527-530.

Muegge, B., Kuczynski, J., Knights, D., Clemente, J., Gonzalez, A., Fontana, L., Henrissat, B., Knight, R. & Gordon, J. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332**: 970-974.

Nyholm, S. V. & McFall-Ngai, M. J. 2004. The winnowing: Establishing the squid-Vibrio symbiosis. *Nature Reviews Microbiology* **2**: 632-642.

Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J.-B. N., Peeters, M., Hahn, B. H. & Hugenholtz, P. 2010. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biology* **8**: e1000546.

Okou, D., Steinberg, K., Middle, C., Cutler, D., Albert, T. & Zwick, M. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4**: 907-909.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H. & Wagner, H. (2011) Vegan: Community ecology package. R package version 2.0-2. http://CRAN.R-project.org/package=vegan. pp.

Olsen, K., Henriksen, M., Bisgaard, M., Nielsen, O. L. & Christensen, H. 2008. Investigation of chicken intestinal bacterial communities by 16S rRNA targeted fluorescence in situ hybridization. *Antonie van Leeuwenhoek* **94**: 423-437.

Ortega, C. & Cruz, A. 1992. Differential growth patterns of nestling brown-headed cowbirds and yellow-headed blackbirds. *The Auk* **109**: 368-376.

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. 2007. Development of the human infant intestinal microbiota. *PLoS Biology* **5**: 1556-1573.

Petnicki-Ocwieja, T., Hrncir, T., Liu, Y. J., Biswas, A., Hudcovic, T., Tlaskalova-Hogenova, H. & Kobayashi, K. S. 2009. Nod2 is required for the regulation of commensal microbiota in the intestine. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 15813-15818.

Phillips, C. D., Phelan, G., Dowd, S. E., Mcdonough, M. M., Ferguson, A. W., Delton Hanson, J., Siles, L., Ordóñez-Garza, N., San Francisco, M. & Baker, R. J. 2012. Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Molecular Ecology* **21**: 2617-2627.

Price, M. N., Dehal, P. S. & Arkin, A. P. 2009. FastTree: Computing large minimum-evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* **26**: 1641-1650.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J., Hansen, T., Le, P., D, Linneberg, A., Nielsen, H., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., MetaHIT, C., Bork, P., Ehrlich, S. & Wang, J. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59-65.

Qu, A., Brulc, J. M., Wilson, M. K., Law, B. F., Theoret, J. R., Joens, L. A., Konkel, M. E., Angly, F., Dinsdale, E. A. & Edwards, R. A. 2008. Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. *PLoS One* **3**: e2945.

Quinn, G. & Keough, M. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, United Kingdom.

R Development Core Team (2010) R: a language for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rajilić-Stojanović, M., Smidt, H. & De Vos, W. M. 2007. Diversity of the human gastrointestinal tract microbiota revisited. *Environmental Microbiology* **9**: 2125-2136.

Rawls, J. F., Mahowald, M. A., Ley, R. E. & Gordon, J. I. 2006. Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. *Cell* **127**: 423-433.

Remsen, J. V., Jr., Cadena, C. D., Jaramillo, A., Nores, M. A., Pacheco, J. F., Perez-Eman, J., Robbins, M. B., Stiles, F. G., Stotz, D. F. & Zimmer, K. J. Version 7 December 2012. A classification of the bird species of South America. *American Ornithologists' Union*: http://www.museum.lsu.edu/~Remsen/SACCBaseline.html.

Schloss, P. D. & Handelsman, J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* **71**: 1501-1506.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. & Weber, C. F. 2009. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* **75**: 7537-7541.

Scupham, A. 2007. Succession in the intestinal microbiota of preadolescent turkeys. *FEMS Microbiology Ecology* **60**: 136-147.

Scupham, A. 2009. Campylobacter colonization of the turkey intestine in the context of microbial community development. *Applied and Environmental Microbiology* **75**: 3564-3571.

Sears, C. 2005. A dynamic partnership: celebrating our gut flora. *Anaerobe* **11**: 247-251.

Sekirov, I., Russell, S. L., Antunes, L. C. M. & Finlay, B. B. 2010. Gut microbiota in health and disease. *Physiological Reviews* **90**: 859-904.

Sharon, G., Segal, D., Ringo, J. M., Hefetz, A., Zilber-Rosenberg, I. & Rosenberg, E. 2010. Commensal bacteria play a role in mating preference of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 20051-20056.

Simpson, J. M., McCracken, V. J., Gaskins, H. R. & Mackie, R. I. 2000. Denaturing gradient gel electrophoresis analysis of 16S ribosomal DNA amplicons to monitor changes in fecal

bacterial populations of weaning pigs after introduction of Lactobacillus reuteri strain MM53. *Applied and Environmental Microbiology* **66**: 4705-4714.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M. & Birol, İ. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research* **19**: 1117-1123.

Slater, G. & Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Smillie, C. S., Smith, M. B., Friedman, J., Cordero, O. X., David, L. A. & Alm, E. J. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**: 241-4.

Spaw, C. D. & Rohwer, S. 1987. A comparative study of eggshell thickness in cowbirds and other passerines. *The Condor* **89**: 307-318.

Stainier, D. 2005. No organ left behind: tales of gut development and evolution. *Science* **307**: 1902-1904.

Stevens, C. E. & Hume, I. D. 1995. *Comparative Physiology of the Vertebrate Digestive System*. Cambridge University Press, Cambridge, UK.

Stewart, J. A., Chadwick, V. S. & Murray, A. 2005. Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *Journal of Medical Microbiology* **54**: 1239-1242.

Thornton, K. 2003. libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* **19**: 3.

Tims, S., Zoetendal, E. G., De Vos, W. M. & Kleerebezem, M. (2011) Host Genotype and the Effect on Microbial Communities. In: *Metagenomics of the Human Body*, (Nelson, K. E., ed.). pp. 15-41. Springer New York.

Turnbaugh, P. J., Quince, C., Faith, J. J., Mchardy, A. C., Yatsunenko, T., Niazi, F., Affourtit, J., Egholm, M., Henrissat, B., Knight, R. & Gordon, J. I. 2010. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 7503-7508.

Vaishampayan, P. A., Kuehl, J. V., Froula, J. L., Morgan, J. L., Ochman, H. & Francino, M. P. 2010. Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome Biology and Evolution* **2**: 53-66.

Van de Merwe, J. P., Stegeman, J. H. & Hazenberg, M. P. 1983. The resident faecal flora is determined by genetic characteristics of the host. Implications for Crohn's disease? *Antonie van Leeuwenhoek* **49**: 119-124.

van Orsouw, N., Hogers, R., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., van der Poel, H., van Oeveren, J. & Verstegen, H. 2007. Complexity Reduction of Polymorphic Sequences (CRoPS): A novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One* **2**: 1172.

Vispo, C. & Karasov, W. H. 1996. The interaction of avian gut microbes and their host: An elusive symbiosis. *Gastrointestinal Microbiology: Gastrointestinal Ecosystems and Fermentations* **1**.

Walter, J. & Ley, R. 2011. The Human Gut Microbiome: Ecology and recent evolutionary changes. *Annual Review of Microbiology* **65**: 411-429.

Wang, Q., Garrity, G., Tiedje, J. M. & Cole, J. R. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**: 7.

Whiteman, N. K., Kimball, R. T. & Parker, P. G. 2007. Co-phylogeography and comparative population genetics of the threatened Galápagos hawk and three ectoparasite species: ecology shapes population histories within parasite communities. *Molecular Ecology* **16**: 4759-4773.

Whitman, W., Coleman, D. & Wiebe, W. 1998. Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95**: 6578-6583.

Wienemann, T., Schmitt-Wagner, D., Meuser, K., Segelbacher, G., Schink, B., Brune, A. & Berthold, P. 2011. The bacterial microbiota in the ceca of Capercaillie (*Tetrao urogallus*) differs between wild and captive birds. *Systematic and Applied Microbiology* **34**: 542-551.

Woolfenden, B. E., Gibbs, H. L., Sealy, S. G. & McMaster, D. G. 2003. Host use and fecundity of individual female brown-headed cowbirds. *Animal behaviour* **66**: 95-106.

Xu, J., Bjursell, M. K., Himrod, J., Deng, S., Carmichael, L. K., Chiang, H. C., Hooper, L. V. & Gordon, J. I. 2003. A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science* **299**: 2074-6.

Xu, J. & Gordon, J. I. 2003. Honor thy symbionts. *Proceedings of the National Academy of Sciences* **100**: 10452-10459.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R. & Gordon, J. I. 2012. Human gut microbiome viewed across age and geography. *Nature* **486**: 222-227.

Yin, Y., Lei, F., Zhu, L., Li, S., Wu, Z., Zhang, R., Gao, G. F., Zhu, B. & Wang, X. 2009. Exposure of different bacterial inocula to newborn chicken affects gut microbiota development and ileum gene expression. *The ISME Journal* **4**: 367-376.

Yoder, J. B., Smith, C. I. & Pellmyr, O. 2010. How to become a yucca moth: Minimal trait evolution needed to establish the obligate pollination mutualism. *Biological journal of the Linnean Society Linnean Society of London* **100**: 847-855.

Zellmer, A. J., Koopman, M. M., Hird, S. M. & Carstens, B. C. 2012. Deep phylogeographic structure and environmental differentiation in the carnivorous plant *Sarracenia alata*. *Systematic Biology* **61**: 763-777.

Zerbino, D. R. & Birney, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. . *Genome Research* **18**: 821-829.

Zhang, H., DiBaise, J. K., Zuccolo, A., Kudrna, D., Braidotti, M., Yu, Y., Parameswaran, P., Crowell, M. D., Wing, R., Rittmann, B. E. & Krajmalnik-Brown, R. 2009. Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences*.

Zoetendal, E., Collier, C., Koike, S., Mackie, R. I. & Gaskins, H. 2004. Molecular ecological analysis of the gastrointestinal microbiota: A review. *The Journal of Nutrition* **134**: 465-472.

Zoetendal, E. G., Akkermans, A. D. L., Akkermans-van Vliet, W. M., de Visser, J. A. G. M. & de Vos, W. M. 2001. The host genotype affects the bacterial community in the human gastronintestinal tract. *Microbial Ecology in Health and Disease* **13**: 129-134.

Appendix A.
PRGMATIC README

PRGMATIC [v.1] README

1. INSTALLATION
The pipeline is dependent on several other pieces of software. These are in the DependentSoftware folder. The packages have been included with PRGMATIC, but if you have difficulties installing them, please see the original webpages. The pipeline was written on MacOSX and requires that the Developer Tools be installed on your machine. These are on the supplementary disc that comes with Macs.

2. EASY INSTALL
1. Open a terminal window
2. cd into the DependentSoftware folder.
3. Type “chmod 755 setup.pl”
4. Type “./setup.pl”.

You should see some output to the screen and when the script has finished, there should be three executables (bwa, cap3, samtools) and a script (samtools.pl) in the PRGMATIC folder.

3. MORE DETAILED INSTALL:

a. BWA:

Included is MacOSX version, which should also work on some other platforms as well. To install, double click bwa-0.5.8a.tar.bz2. Open a terminal window and cd into the bwa-0.5.8a folder. Type "make" (without quotation marks). When this finishes, make a copy of the bwa executable and put the copy in the PRGMATIC folder. This is correctly installed if you double click the executable and a screen with a parameter list opens.
http://sourceforge.net/projects/bio-bwa/files/bwa-0.5.8a.tar.bz2/download
http://bio-bwa.sourceforge.net/

b. CAP3:

Included is the 64-bit, MacOSX version of this program. To install CAP3, double-click cap3.macosx.intel64.tar. When it unpacks into a folder, make a copy of the cap3 executable and put the copy in the PRGMATIC folder. This is correctly installed if you double click the executable and a screen with a parameter list opens.
http://seq.cs.iastate.edu/cap3.html.

c. MUSCLE

Muscle is the sequence aligner for after the loci have been called. It is executable from the file included with PRGMATIC and shouldn't require anything to make it run as long as it's in the correct folder. You can double check that it is correctly installed if you type "chmod 755 muscle3.8.31_i86darwin64" then "./muscle3.8.31_i86darwin64" from a terminal window inside the PRGMATIC directory. The included version is for 64-bit Intel MacOSX machines – if you have different hardware or need to get an original copy, see http://www.drive5.com/muscle/.

d. SAMTOOLS:

Included is the MacOSX version, which should also work on some other platforms as well. To install, double click samtools-0.1.8.tar.bz2. Open a terminal window and cd into samtools-0.1.8. Type "make" (without quotation marks). When this finishes, make a copy of the samtools executable and put the copy in the PRGMATIC folder. This is correctly installed if you double click the executable and a screen with a parameter list opens. Also make a copy of samtools.pl from the misc folder and put that copy in the PRGMATIC folder. (samtools.pl is not an executable, so as long as it is in the correct folder, it is correctly "installed".)
http://sourceforge.net/projects/samtools/

e. VARSCAN

The included VarScan should already be installed upon downloading the PRGMATIC package. You can check this by opening a terminal, cd into PRGMATIC and type "java –jar VarScan.v2.2.2.jar". This is correctly installed if you see a screen with a parameter list.
http://varscan.sourceforge.net/

f. COMPUTE
The compute package is more difficult to install than the other software. I recommend following the instructions at http://molpopgen.org/software/libsequence.html to get the necessary library (libsequence) and the instructions at http://molpopgen.org/software/lseqsoftware.html to get the analysis package (which contains compute). Once you have this installed, make a copy of the compute executable and put it in the PRGMATIC folder. This is optional software, so if it is not installed on your machine, PRGMATIC will still run, you just need to turn off the part of PRGMATIC.pl that calls compute.

4. TO TURN OFF THE COMPUTE ANALYSIS OR MUSCLE ALIGNMENT
Open PRGMATIC.pl in a text editor. On line 74, it should say "multiCompute();". Put a pound sign (#) in front of this line and save the file. That should effectively turn off calling the compute package and there should be no errors. To turn off the MUSCLE multi-sequence alignment, open the PRGMATIC.pl in a text editor and put a pound sign (#) in from of line
73, which should read "muscleAlignments()". Save the file.
5. PRGMATIC FOLDER
In the PRGMATIC folder you should now have 11 things:

1 BWA executable
2 calledAlleles folder
3 CAP3 executable
4 COMPUTE executable (optional)
5 DependentSoftware folder
6 inputFASTA folder (with alignedLoci folder inside)
7 MUSCLE3.8.31_I86DARWIN64
8 PRGMATIC.pl
9 SAMTOOLS executable
10 SAMTOOLS.PL
11 VARSCAN.V2.2.2.JAR

Inside the inputFASTA folder you should place your tag separated fasta files.

6. (MY PREFERRED METHOD FOR) GENERATING FASTA FILES – USING RDP WEBSITE
Off the 454 machine, you should have gotten at least one .fna file, one .qual file and a folder of .sff files. To quality control the reads, I use the Ribosomal Database Project's Pyrosequencing Pipeline (at http://pyro.cme.msu.edu/). Their "Pipeline Initial Process" is easy to use, fast and on its own server, so there's nothing to download.

RDP Pyrosequencing Pipeline Initial Process Parameters:
Sequence file in FASTA format: (upload the .fna file here)
Quality file in FASTA format (optional): (upload the .qual file here)
Upload a tag file: (upload your tag file here – this is the file that says which tag sequence belongs to which individual. The format is very easy; on a new line for each individual: TagSequence (tab) IndividualName. **When you run the bigFastaRename.pl script to rename the files, the tag file will need to have UNIX line breaks.**)

Gene name: Other
Forward Primers: (paste your forward primer here)
Reverse Primers: (paste your reverse primer here)
FILTERS
Forward primer max edit distance (0 to 2): (this refers to how many errors you'll allow in your forward primer sequence. I use 2 but if you're being conservative, 0 or 1 will weed out more sequences.)
Reverse primer max edit distance (0 to 2): (same as above but for reverse primer. Again, I use 2.)
Max number of N's: 0 (I highly recommend using 0 here, since one ambiguous base can be indicative of error prone sequence. See Huse et al. (2007) for a good overview of 454 generated errors).
Min sequence length (>=50): (I use either 100 or 150, depending on the dataset, but you can use whatever you deem appropriate. Shorter sequences are more error prone, but throwing out reads unnecessarily is suboptimal).
Minimum Average Exp Quality Score: (20 is what I usually use. This corresponds to an average error rate of no more than 1/100 bases having an error (DOUBLE CHECK THIS). Higher number here results in fewer reads passing the filter with a higher quality score.)
Keep primers: (Do not check)

Click "Perform Initial Processing"

This is generally very fast and I have my file downloading within 10 minutes (usually). It may take longer for big files. Once the file downloads, double click it and it will unpack a folder with your sequences separated by the names in the tag file. The files in these folders named "Individual_trimmed.fasta" and "Individual_trimmed.qual" can be used as input for the pipeline. I've included a script in the DependentSoftware folder ("bigFastaRename.pl") that renames the sequences in these folders (and their associated quality scores) as IndividualName_00001, IndividualName_00002, etc. and puts them in files called IndividualName.fasta and IndividualName.qual. Renaming the sequences makes them easier to view and understand later, when knowing which reads came from which individual is helpful. I highly recommend running bigFastaRename.pl but it is not necessary.

7. TO RUN BIGFASTARENAME.PL:
1. Copy or drag the file into the RDP downloaded folder.
2. Open a terminal window and cd into this folder.
3. Type "chmod 755 bigFastaRename.pl" to give yourself permission to run the script.
4. Type "./bigFastaRename.pl".
5. You should see the prompt "Drag tagfile here:" on the screen. Drag and drop the tagfile there (it doesn't need to be in the same folder as everything else) and the script should run through all the folders in the tagfile, outputting a .fasta and a .qual file for each one. IF THIS DOES NOT HAPPEN – check the line breaks on your tag file. They should be Unix or the script will just read the first line of the tagfile and quit.
6. Copy or drag these files into the "inputFASTA" folder in the PRGMATIC folder. You're ready to go!

*****If the script runs for the first individual but stops after that, the line breaks in the tag file need to be changed to UNIX. The program is reading the first line then hitting a hard return it doesn't understand and stopping. Changing the line breaks will fix this. *****

8. TO RUN PRGMATIC:
Make sure your individual .fasta files are in the inputFASTA folder (quality files are optional, but should be in this folder if you have them). Also, the quality files need to be of the "same name" as the .fasta file they're associated with (e.g. IND01.fasta should have a quality file named IND01.qual – this is a requirement of cap3).
Make sure you have the 11 things listed above in working order in the PRGMATIC folder.
Open a terminal window. cd to the PRGMATIC folder. Type "chmod 755 PRGMATIC.pl" to give yourself permission to run the script. Type "./PRGMATIC.pl" to run the script.

9. INPUT PARAMETERS:
When the program runs, you'll see a variety of prompts and you need to give the script some information.

"Enter the dataset nickname: " (don't use an underscore in the nickname _)
What you enter here doesn't have an effect on how the program runs, it's just a way of giving a name to the various output files that the pipeline generates. I generally use something informative and short. Like Trial0915 for a test run that I did on 15 September. It could also be the focal taxa or locality or whatever.

Parameter Settings
"Minimum number of reads to call an allele (5):"
To generate the pseudo-reference genome, the pipeline calls "high confidence alleles" from clusters within an individual. This parameter sets the minimum number of reads you want for a cluster to be called an allele. If you'd like to use the default, type 5. If you'd like a more conservative p-rg, enter a higher number, if you'd like to be more liberal, enter a lower number.

"Minimum % identity to call a locus (90):"
To generate the p-rg, the pipeline clusters the high-confidence alleles at a given percent identity (similarity). 90% seems to work pretty well.

"Minimum coverage for calling consensus in an individual (3):"
Once all the reads have been blasted to the p-rg, VarScan calls a consensus sequence for each locus. Here you can set the minimum number of reads you'd like to call the consensus sequence. Higher values = more conservative. Lower values = more liberal.

"Minimum coverage for calling a SNP (3):"
VarScan also calls SNPs from the reads that blast to each locus. Here you can set the minimum number of reads that support a SNP for it to be called an actual variant. I wouldn't go below 2, since you'd then be calling every SNP (aka every error) a real variant.

“Minimum % of reads for calling a SNP (20):”
In addition to a minimum raw number of reads being necessary for a SNP to be called, those reads must also represent a certain percentage of the coverage at that base. 20% seems pretty standard in the literature. Higher numbers will be more conservative and lower numbers will be more liberal.

10. OUTPUT
When the pipeline is finished running, there should be .fasta files in the calledAlleles folder corresponding to the loci in the p-rg. These should be viewable in any program that reads .fasta files.
The “.counts.txt” file displays how many individuals were called for each locus. This can be opened in Excel and then sorted to show which loci contained all or most individuals (i.e. those of highest interest).
The “.HoHe.txt” file contains how many individuals, heterozygotes and alleles were called for each locus as well as the observed and expected heterozygosities. This file can also be opened directly by Excel.
The “.ComputeOutput.txt” contains the output from the compute analyses. This can also be opened in Excel.
The “MultiHitLoci.txt” contains information on where there are more than two base pairs for a given position on the reference genome within an individual. This is highly suspect, especially if more than two base pairs occur at high frequency (i.e., more than one of a given base). If more than one individual appears on the list for a single locus, that locus is almost definitely paralogous. It might even be wise to throw out every locus on the list, or rerun the analysis with higher cutoffs for calling a locus.

11. TEST DATA
To run the test data, copy the three .fasta and three .qual files into the inputFASTA folder. Run PRGMATIC.pl. There should be a lot of output to the screen. In the zip file with the test data, I put the .counts.txt, .HoHe.txt, and .ComputeOutput.txt files that were generated on my machine. They should match the files you get.

12. CONTACT
Please feel free to contact me about any issues you’re having with PRGMATIC or the dependent software. I’d be more than happy to do what I can –
Sarah Hird
shird1@tigers.lsu.edu

13. RECOMMENDED READING
Huse SM, Huber JA, Morrison HG, Sogin ML and DM Welch. 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8: R143 (doi: 10.1186/gb-2007-8-7-r143)

14. CITATIONS
If you use PRGMATIC, please cite (all of) the following papers:

HIRD, S.M., BRUMFIELD, R.T. & CARSTENS, B.C. 2011. PRGMATIC: an efficient pipeline for collating genome-reduced second generations sequencing data using a pseudo reference genome. Molecular Ecology Resources, 11:743-748.
EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32, 6.
HUANG, X. & MADAN, A. 1999. CAP3: A DNA sequence assembly program. Genome Research, 9, 10.
KOBOLDT, D., CHEN, K., WYLIE, T., LARSON, D., MCLELLAN, M., MARDIS, E. R., WEINSTOCK, G., WILSON, R. K. & DING, L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics, 25, 3.
LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 7.
LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GROUP, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2.

Optional program citations:
Compute:
THORNTON, K. 2003. libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics, 19, 3.
RDP:
COLE, J. R., WANG, Q., CARDENAS, E., FISH, J., CHAI, B., FARRIS, R. J., KULAM-SYED-MOHIDEEN, A. S., MCGARRELL, D. M., MARSH, T., GARRITY, G. M. & TIEDJE, J. M. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research, 37, 5.
Tablet:
MILNE, I., BAYER, M., CARDLE, L., SHAW, P., STEPHEN, G., WRIGHT, F. & MARSHALL, D. 2009. Tablet - next generation sequence assembly visualization. Bioinformatics, 26, 2.

Updated: 7 January 2011 by S. Hird

Appendix B.
PRGMATIC: Guide To Common Errors

Next-gen sequencing data is rife with errors, as it can (and does) come from many sources: PCR error, sequencing error (homopolymer miscalls, indels…) and analysis. PRGMATIC is not intended to be a "black box" for analysis – the user must look at the data and determine that the calls are correct. This guide is to show you what some of these errors look like in order for the user to be more confident about calling a locus good or bad and about reanalyzing data or adjusting parameters as needed.

PRGMATIC outputs a variety of files that allow you to view both the clusters within an individual (the "alleles") and the reads within an individual blasted to the PRG. The alleles are in the inputFASTA folder and have the suffix .cap.ace (example IND01.fasta.cap.ace – if IND01.fasta is the fasta file of the original reads). If you have installed the program Tablet, you can double click the .ace files and they should open. Alleles should be almost identical and therefore don't need to be scrutinized by the user all that much.

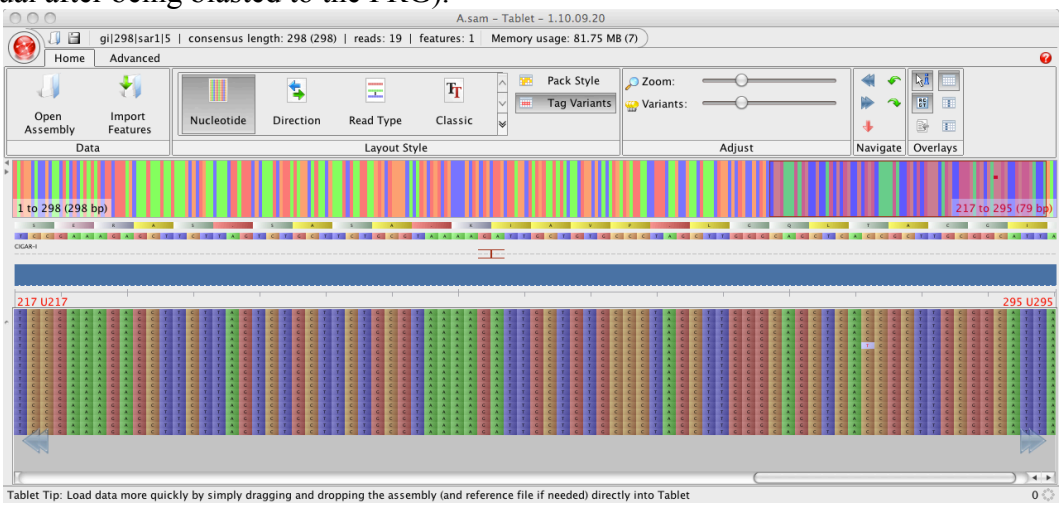
The alleles are then grouped into loci. You may look at the reads that comprise a locus by opening the file in the PRGMATIC folder that has "total_" as a prefix, followed by your nickname and ending with .cap.ace as the suffix (example, if your nickname is trial, the file is total_trial.fasta.cap.ace).

Finally, you may view the alignments of all the reads within an individual to the PRG. **These files are the most important to view and confirm that the reads comprise a believable homologous locus. You should look at many of these files across individuals and loci and any locus that looks suspect based on the summary files output by PRGMATIC.** These must be imported to Tablet using the .sam file as input and the .fna file as the reference genome. (Each individual has a .sam file but there is only one reference genome (.fna file) for a particular run. It will not change across individuals.)

The following show some typical outputs of PRGMATIC and should give you an idea of what common errors look like (and what good loci look like too). Please feel free to contact me with questions –

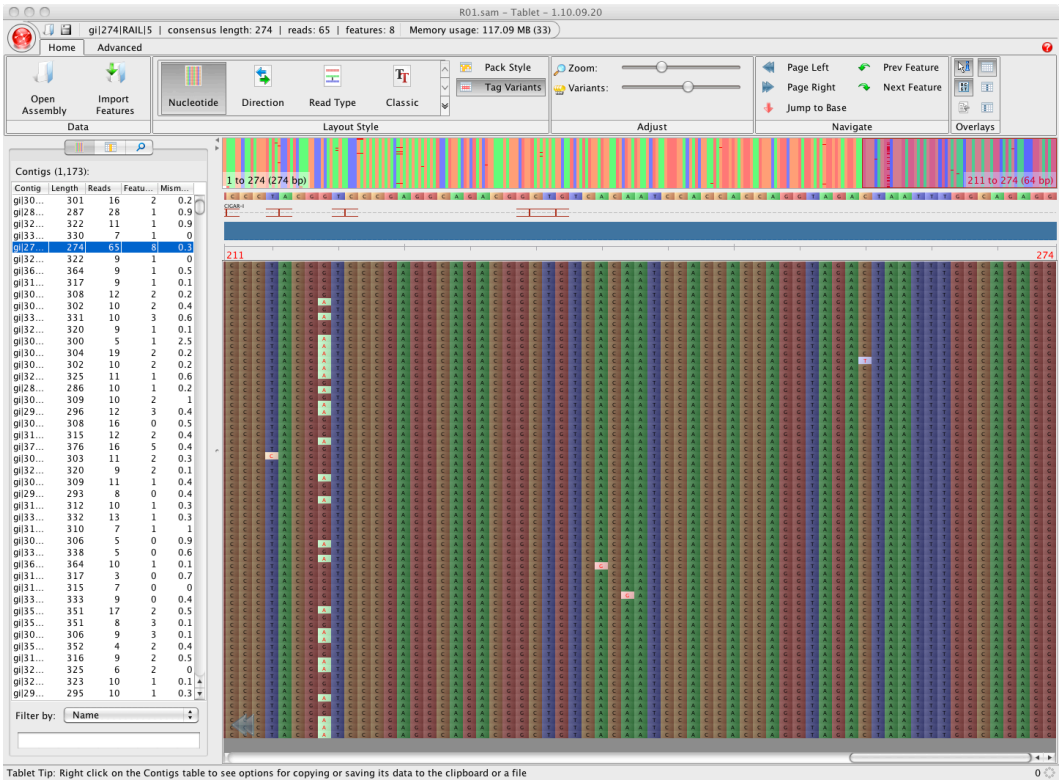
Sarah Hird (sarah.hird@gmail.com)

1. A good monomorphic locus (this – and all subsequent examples - is a .sam file of an individual after being blasted to the PRG).

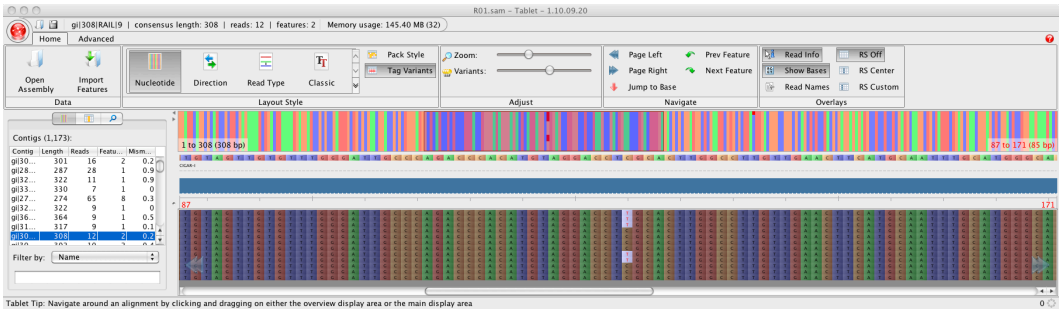


Notice that there are very few SNPs and the only one we can see is a single base from a single read. I feel we can be fairly confident that this is an error, as there are 18 reads that support the consensus base and only the one supporting the SNP. I refer to these as "singleton errors" and PRGMATIC should ignore these.

2. Two good polymorphic loci.
2A:



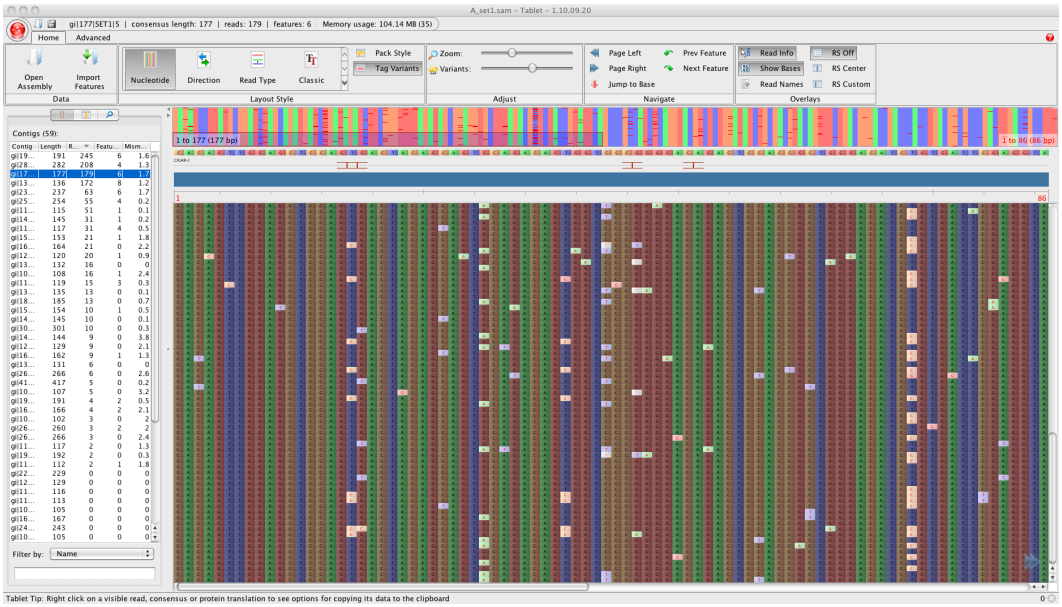
2B:



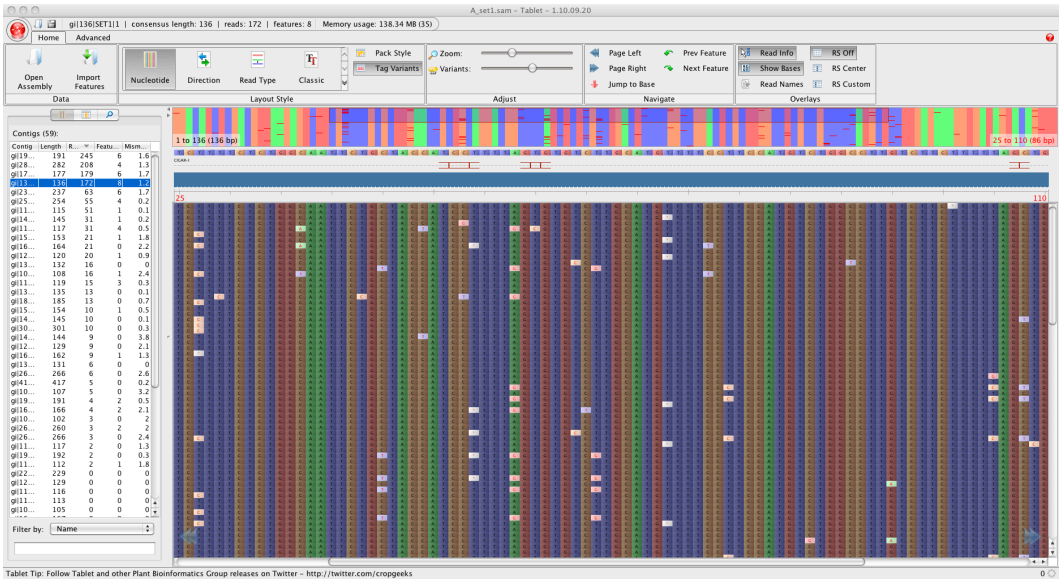
Here we have genuine SNPs. In 2A, at base 219, there are 25 reads with an A and 40 reads with a G. In 2B, there are 5 T's and 7 C's. In a perfect world, all heterozygosity would show up as exactly 50% of the reads – how close to that you want to be is up to you. There are still singleton errors in 2A, but they are unsupported by more than a single base so I feel confident those are errors.

3. Bad (possibly paralogous) loci.

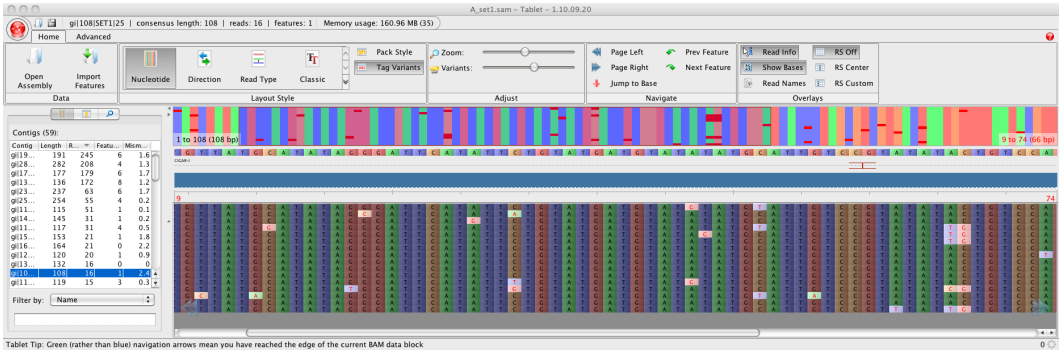
3A:



3B:

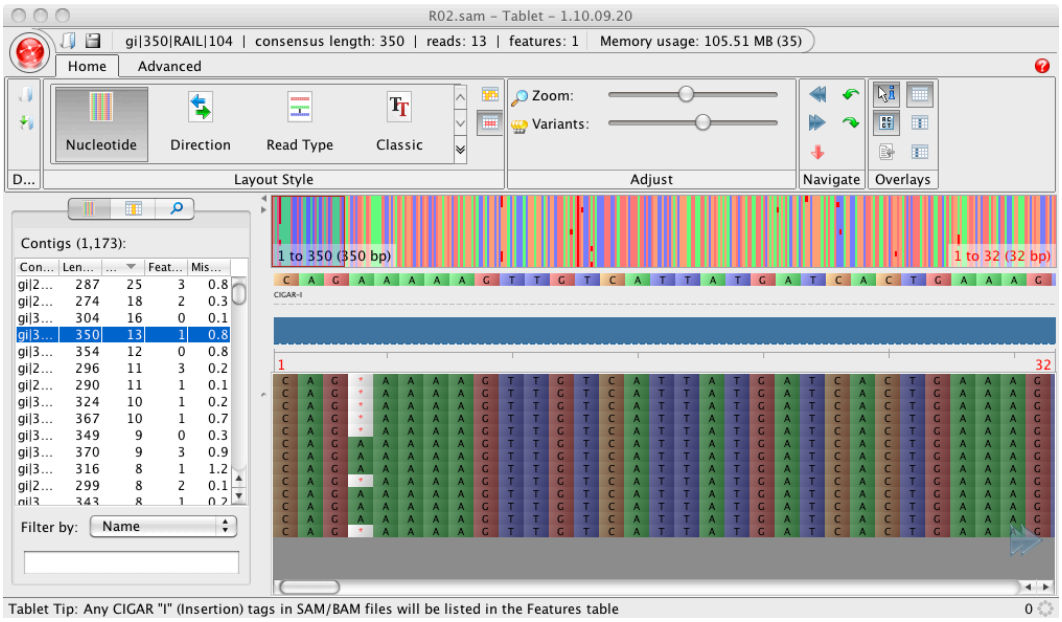


3C:



When you see a “checkerboard” like this, it is probably obvious that something is wrong and we should not use this locus. There are SNPs everywhere and we cannot resolve two alleles within this individual. See how there are several bases that have a significant number of SNPs (i.e., COLUMNS of data that have multiples of the same SNP)? If these were reads from a single homologous locus within a diploid individual, we would not see this much (and this pattern of) variation. So what is going on here? There are several possibilities, but when you can see multiple possible unique alleles (i.e., >2), I think the most parsimonious explanation may be paralogous loci. The other possibility is extreme PCR error. This pattern is prone to showing up in the “loci” that are called within most/all individuals. **Please check all loci that contain all individuals for obvious paralogy!** Also notice that in the middle of 3C there is a column with all 4 base pairs present – a good sign that something is wrong! When more than 2 bases show up in a position, treat the locus with suspicion – a diploid individual can’t have more than 2 for a single locus.

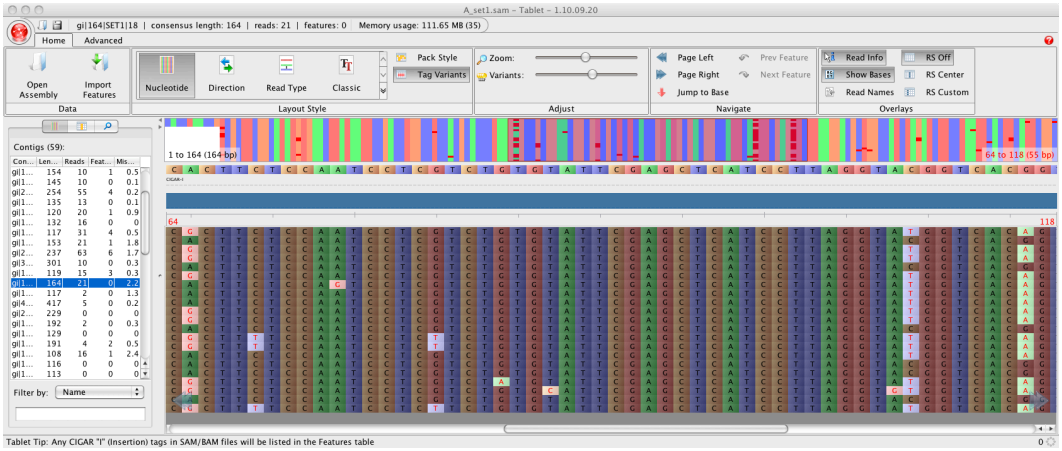
4. Indel errors



The most common types of error with 454 data are associated with homopolymers. First, the machine may have trouble telling how many A's are in a row once there have been several consecutively but there is also an error that causes "SNPs" close to the homopolymer. This is not to say that all gaps are errors, but it is particularly difficult to resolve these, I think. Here we have 7 out of 13 reads supporting a gap (and 6 reads supporting an A). Is this a legitimate SNP? Seeing as the SNP is at the beginning of a string of A's, I'd be inclined to err on the conservative side and call this as a homozygote for 4 A's. However, it's entirely possible this is an indel SNP. Perhaps looking at this locus in other individuals will help resolve the issue, but since an identical error is probable in every case of a homopolymer, it's still hard to say.

5. Tricky loci

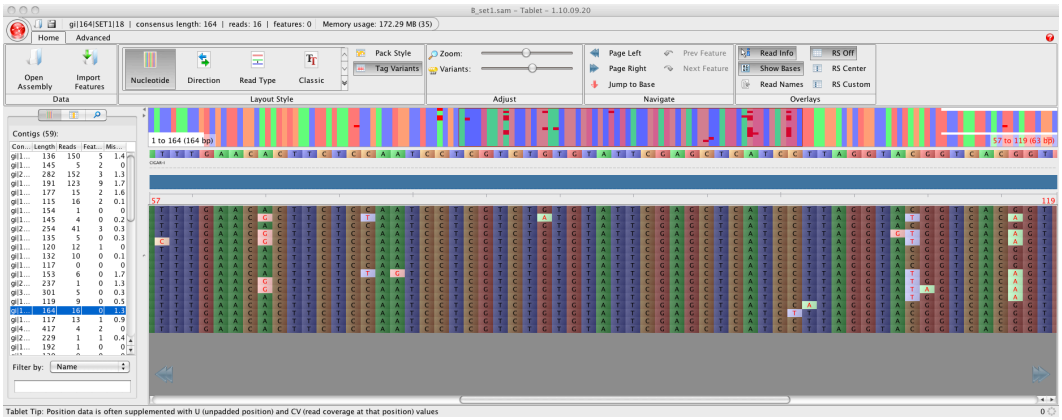
5A:



Sometimes loci are really, really difficult to determine. This looks like it could be an extremely variable locus within an individual. The three main columns of SNPs look like they might be resolvable to 2 alleles. However, if we look closely, the three SNPs don't partition out to exactly 2 alleles (i.e., not every read that has an A at position 65 has a C at 109 and a G at 116). Additionally, there are 2 columns with 3 T SNPs in them (positions 69 and 80). Sure, it's only 3

out of 21 reads, but they are consistent across reads (i.e., every read that has a T at 69 also has a T at 80 AND a G at 65, a T at 109 and an A at 116). So how many alleles do we have here? Again, it might be useful to look at another individual:

5B:



Here is a second individual at the same locus. We still have the 3 main SNPs, but the two minor T SNPs (at 69 and 80) do not appear. Can we resolve 2 alleles in this individual? Still difficult, as there are 2 reads that have a (consensus) A at the first SNP (position 65) but a T at 109 and an A at 116. I suppose this is a matter of personal opinion – I would continue to check the rest of the individuals and probably try to be conservative. Errors do happen but the same error showing up across individuals is less likely than singleton errors.

Appendix C.
LOCINGS README

LOCINGS [v.1] README

LOCINGS: a simple database for reformatting and displaying multi-locus datasets
---

---

1. INSTALLATION

The program is dependent on several other pieces of software. LOCINGS was written on MacOSX. LOCINGS is available for download at https://github.com/SHird/LOCINGS. You will have to install MongoDB separately from the rest of the package.

1.1 MongoDB

1.1.1 Installation of MDB

Go to www.mongodb.org/downloads;
Download correct version (if you don't know if your machine is 32 or 64 bit, see section 1.1.3.);
Double click the downloaded file this should unpack it into a folder called something like "mongodb-osx-x86_64-1.8.2" (I would rename this folder "MongoDB" but you don't have to);
Move the MongoDB folder to Applications (or wherever you want to keep it. ***You will need mongod running every time you use LOCINGS, so you should remember where you put the MongoDB folder.);
Make the directory that stores the data by opening a terminal and typing "mkdir /data/db" (without the quotations)
Go to the mongoDB folder, then the bin folder.
Double click mongod. This should open a screen with something like this:

```
Last login: Wed Aug  3 08:12:53 on ttys001
/Applications/mongodb-osx-x86_64-1.8.1/bin/mongod ; exit;
HappyPappy:~ shird$ /Applications/mongodb-osx-x86_64-1.8.1/bin/mongod ; exit;
/Applications/mongodb-osx-x86_64-1.8.1/bin/mongod --help for help and startup options
Wed Aug  3 09:13:53 [initandlisten] MongoDB starting : pid=11886 port=27017
dbpath=/data/db/ 64-bit
Wed Aug  3 09:13:53 [initandlisten] db version v1.8.1, pdfile version 4.5
Wed Aug  3 09:13:53 [initandlisten] git version: a429cd4f535b2499cc4130b06ff7c26f41c00f04
Wed Aug  3 09:13:53 [initandlisten] build sys info: Darwin erh2.10gen.cc 9.6.0 Darwin Kernel
Version 9.6.0: Mon Nov 24 17:37:00 PST 2008; root:xnu-1228.9.59~1/RELEASE_I386 i386
BOOST_LIB_VERSION=1_40
Wed Aug  3 09:13:53 [initandlisten] waiting for connections on port 27017
Wed Aug  3 09:13:53 [websvr] web admin interface listening on port 28017
```

Leave this window open as you run LOCINGS.

1.1.2. Correctly shutting down MongoDB

When you are finished with LOCINGS and MongoDB, you must shut down MongoDB correctly. Click on the terminal window then press Control + C (the control button and the "c" button at once). If you don't do this, you'll have to find a file and delete it before you can get MongoDB running again, spotlight the file "mongod.lock" then drag it to the trash.

1.1.3. How to tell if your computer is 32- or 64-bit (from support.apple.com/kb/ht3696)
1. Choose About This Mac from the Apple menu in the upper-left menu bar, then click More Info.
2. Open the Hardware section.
3. Locate the Processor Name [under the Hardware Overview].
4. Compare your Processor Name to information below to determine whether your Mac has a 32-bit or 64-bit processor.

Processor Name	32- or 64-bit
Intel Core Solo	32 bit
Intel Core Duo	32 bit
Intel Core 2 Duo	64 bit
Intel Quad-Core Xeon	64 bit
Dual-Core Intel Xeon	64 bit
Quad-Core Intel Xeon	64 bit
Core i3	64 bit
Core i5	64 bit
Core i7	64 bit

1.2. LOCINGS Easy Install
Download the LOCINGS package from GitHub. You may either go to the website [https://github.com/SHird/LOCINGS] and click the "Downloads" button (then choose the .tar.gz option). When the package has downloaded, double click the download, rename it to "LOCINGS" and move the folder to the Applications folder (or wherever you'd like it to be).
Alternatively, if you have git on your machine, you can clone the directory by typing "git clone git@github.com:SHird/LOCINGS.git /Applications/LOCINGS". (This will install LOCINGS into the Applications folder on MacOSX. You may move it.)

Open a terminal window, cd into the LOCINGS folder
Type "python setup.py install" (each of these steps may take a couple of minutes to complete and will print lots of output to the terminal window)
Type "python distribute_setup.py"
Type "easy_install numpy" (lots of ugly output)
Type "easy_install biopython"

There should be a lot of output to the terminal screen.
Once the last command has finished, type "run_LOCINGS.py" - this should open a new window with a short greeting and instructions...

1.3 More information
1.3.1 Python installation
MacOSX should come with a version of Python already installed. If you're not sure Python is on your machine:
In an open terminal, type "python". This should display something like "Python 2.7.1 (r271:86882M, Nov 30 2010, 10:35:34)" plus a couple of other lines of text followed by ">>>". Type "exit()".
If the above did not happen, go to www.python.org/getit/ for Python installation.

1.3.2 NumPy - http://sourceforge.net/projects/numpy/files/ (they have an automated installer, download the file, double click, follow instructions)
For more information: http://www.scipy.org/Download. NumPy MUST be installed BEFORE Biopython!

1.3.3 Biopython - http://biopython.org/wiki/Download; download, double click, move the folder into the home directory, open terminal window, cd into folder, type three commands: [1] "python setup.py build", [2] "python setup.py test", [3] "sudo python setup.py install"

1.3.4 PyMongo - https://github.com/mongodb/mongo-python-driver; click "Downloads" button (upper-ish right of screen); click "1.11"; once it downloads, rename the folder "pymongo", put the folder in the home directory, cd into it, and type "python setup.py install"
For more information, other installation options:
http://api.mongodb.org/python/current/installation.html

1.3.5 pysam - http://code.google.com/p/pysam/downloads/list; download file, double-click, move folder into Applications folder (or where you'd like to keep it), cd into folder and type two commands: [1] "python setup.py build", [2] "python setup.py install"

1.3.6 seqlite_mod.py, ez_setup.py & distribute_setup.py, these module are included with the distribution and you should not have to do much to install them.

2. TO RUN LOCINGS:
2.1. Short answer: find and double click "mongod"; type "run_LOCINGS.py" in a terminal window

2.2. Contents of the folder
I think it's easiest to keep everything in one place, but you don't necessarily have to. You need the following components in the LOCINGS folder in order to run it:

1. locings folder
2. scripts folder
3. setup.py
4. README.txt
5. locus folder of fasta files*
6. BAM and BAI folder*
7. Tab-delimited text file of demographic data*
8. IMa2 input file (if formatting for IMa2 or Migrate, format described below)

*These folders/files don't need to be in the LOCINGS folder.
After the program is run for the first time, an additional four files will appear with the extension .pyc (for each of the python scripts above). These can be removed but are used to speed up the running of python scripts.

2.3. Starting MongoDB
MongoDB must be running before LOCINGS is started. To do this, you can find "mongod" with the Finder and double click it; or go to the MongoDB/bin folder and double click "mongod"; or, from the terminal, you can cd into the MongoDB folder, then cd into "bin" and type "./mongod". Leave the window open and when you are finished with LOCINGS YOU MUST EXIT MONGOD BY PRESSING CONTROL+C (TOGETHER). If you forget to do this, the mongod process will be improperly shut down and not start the next time you need it. If that happens, see Section 1.1.2. above.

2.4 Starting LOCINGS
To run the program, open a terminal window and type "run_LOCINGS.py". This should open a new window with these basic instructions: "Please enter the data in the order listed in the Import Menu. Once data has been loaded via the Import Menu, press ,ÄòDisplay the data'." Don't close this window ,Äì it will close the program.

3. INPUT FORMATS
3.1 Import types
LOCINGS uses three imports.
[1] The first is a folder of files that contain loci in fasta format. These files should have ".fasta" as their final extension. The folder may be located anywhere.
[2] The second import is a folder of indexed bam files from a short read aligner. Each bam (and corresponding .bai) file should correspond to an individual in the dataset. I'm working on getting sam format to work too, but for right now, indexed bam works best.
[3] The third is a tab-delimited text file that contains demographic data for the individuals in the dataset. There must be at least two columns, labeled "Individual" and "Population", which contain information on the name of the individual (as it appears in locus files and BAM file names) and which population the individual came from (these can be numbers or letters). The file may contain as many columns as you'd like, they will appear on the "Summary Screen" of the program.
*LOCINGS will run and reformat loci if just [1] loci and [3] demographic data are entered.

3.2 Locus Names **IMPORTANT**
The names of the files from import [1] that correspond to loci need to match with the locus names in the BAM files, import [2]. Basically, if your loci are called "locus1", "locus2" and "locus3", the fasta files need to be called "locus1.fasta", "locus2.fasta" and "locus3.fasta" and the loci in your BAM file need to be "locus1", "locus2" and "locus3". There are a few exceptions - the locus file name comes from anything before the first "." in the locus file name - so the files could be called "locus1.aln.fasta" or "locus1.080911.fasta" as long as the prefix before the first "." matches the BAM loci. Also, the BAM loci may have prefixes, but this time, due to common GenBank annotation, you may add things to the BAM loci names with a "|" as a separator. So

BAM loci may be called something like "gi|323|testdata|locus1" as long as the last piece of the header matches the fasta loci names.
Any combination of the below files will work in LOCINGS. Please feel free to contact me if you are unsure if your locus names are causing the program difficulty.

FASTA FILE - must contain the locus name before the first period and have ".fasta" as final extension
locus1.fasta
locus1.aln.fasta
locus1.test.data.fasta

BAM LOCUS NAME - must contain the locus name after the final | or as the only text
locus1
fake|any text you want|locus1
gi|323|fakeData|08august2011|locus1

4. IMPORTING DATA
4.1 Loci/fasta file(s)
Step 1 in the import menu will open a window where you should find and select the locus folder. After successful import, the LOCINGS screen will tell you. The terminal window will print data as the files are read. It should be a lot of text that looks something like:
"Got this file: /Users/shird/Desktop/juncoLoci/JUNCOmatic_63_aln.fasta
locusFasta = JUNCOmatic_63_aln.fasta ; individuals {'J12': 0, 'J09': 0, 'J18': 0, 'J19': 0, 'J01': 0, 'J17': 0, 'J03': 0, 'J11': 0, 'J05': 0, 'J04': 0, 'J10': 0, 'J06': 0} ; indInFasta ['J12', 'J09', 'J18', 'J19', 'J01', 'J17', 'J03', 'J11', 'J05', 'J04', 'J10', 'J06'] ; SNPs = 5 ; number alleles = 24 ; length = 284 ; path = /Users/shird/Desktop/juncoLoci/JUNCOmatic_63_aln.fasta"

4.2 SAM/BAM NGS data
Step 2 will import the net-gen alignments - you should find and select the indexed sam or bam folder. LOCINGS will update as the import is finished. The terminal window will print data as the files are read. It will look something like this for each file:
"Got this folder: /Users/shird/Documents/Dropbox/juncoBam
Got this file: /Users/shird/Documents/Dropbox/juncoBam/J01.sorted.bam
730
individuals.J01"
Please note that .bam files are binary versions of .sam files and thus .bam files are smaller and will load faster (perhaps much faster) than .sam files. One way to convert .sam files to .bam files is the "view" tool in the samtools package (http://samtools.sourceforge.net/)

4.1.3 Demographic data
The final step imports the demographic data in a tab delimited file - find and select the tab delimited demographic data file. Again, LOCINGS will tell you when the import has been successful and the terminal will print data for each individual as the files are read:
"{'Longitude': '-109.876', 'Individual': 'J01', 'Location': 'NoPlace, TX', 'Latitude': '45.678', 'Species': 'Junco hyemalis', 'Population': 'POP1'}
POP1"

5. THE LOCINGS INTERFACE

LOCINGS has three screens meant to show you how much data is associated with each individual in your dataset. The first screen will display text updates as the program completes functions. The second is a summary screen where each individual is a row and the demographic data are the columns. There is also a "numLoci" column that displays the number of loci called for that particular individual. If you click one of these numLoci buttons, the third screen appears that displays the specific loci. On this screen there are five columns:

Locus Name = the locus file/locus name

Length = length of the locus

Coverage_This_Ind = how many raw reads from this individual aligned to the locus. If this button is pressed, a fasta file is generated that contains these reads. This file will be printed to the directory that contains the LOCINGS scripts.

Number_Inds = how many individuals are present in the locus (fasta) file

Coverage_Total = how many raw reads from any individual aligned to the locus. If this button is pressed, a fasta file that contains the reads is generated and printed to the LOCINGS directory.

6. EXPORTING DATA

You may export summary data or the raw data in three formats (NEXUS, IMa2 and MIGRATE); you may select either specific individuals from your dataset you'd like to include in your files OR you may select a set of populations. If you select populations, LOCINGS will scan the locus files for any locus that contains at least one individual from each of the selected populations and reformat these loci.

6.1 Summary Data

Selecting the "Export Data -> SummaryData" option will print all the data displayed when you press "Display the data" on the main menu. It will print two files, one for the summary screen and a second that contains all the information for all individuals and loci. This can be a large file but is useful for sorting data in a spreadsheet to report various coverage statistics.

6.2 NEXUS format

NEXUS format requires aligned data. Files will be printed with a ".nex" extension.

6.3 IMa2 format

IMa2 requires an additional input file to describe some of the dataset specific parameters. I've included an example file. Please see IMa2 manual (available at http://lifesci.rutgers.edu/~heylab/ProgramsandData/Programs/IMa2/Using_IMa2_10_13_10.pdf) for more information on the parameters.

The IMa2 extra parameters file must be named "IMa2InputFile.txt" and should follow the format of the included IMa2InputFile format ,Äì just type the values you need after the colon.

Header Line: Type whatever you want here that will identify your IMa run

Population Tree: ((0,1):5,(2,3):4):6

Inheritance Scalar: 1

Mutation Model: H

Mutation Rate (optional): 0.000008

Mutation Rate Range Lower (optional): 0.000007
Mutation Rate Range Upper (optional): 0.000009

If you aren't using the Mutation Rate Parameters, delete them from the file. If you are, please note that the mutation rate you enter needs to be the substitutions / site mutation rate; LOCINGS will multiply this rate by the length of each locus (# of sites) to give a per base mutation rate (which is what the program requires). Also note that the Population Tree is required and is a very specific format, so please check with the IMa2 documentation for how to write it.

Also note, the same inheritance scalar and mutation model will be printed for every locus. If you have different types of data (autosome, mtDNA, X- or Y-linked data), you might want to put the similar types together in a folder and load them independently into LOCINGS, and export with the correct inheritance scalar in the IMa2InputFile. You can then concatenate the files.

6.4 Migrate format
Migrate format requires that an IMa2 additional file be in the folder, but doesn't use the information ‚Äì so if you just need Migrate output, leave the example IMa2InputFile.txt in the folder as is.

7. TEST DATA
I've included a very small test dataset, containing four individuals and five loci.

8. WHAT IF...?
I've attempted compiling a list of potential problems - email me (please) if you encounter something not on this list so I can add it for other users.

8.1. ...LOCINGS won't start?
If this ("pymongo.errors.AutoReconnect: could not find master/primary") is the last line of output printed to the terminal window, it means that mongod is not running. In the MongoDB/bin folder, double click on "mongod" then try starting LOCINGS again.

8.2. ...locus screen coverage buttons all say "0"?
This could be due to the name of the loci in the SAM/BAM files not corresponding correctly to the name of the fasta locus files. Double check by referring to section 3.2 above. Also, if you import the SAM/BAM files before the loci files, you'll get 0s in the coverage columns.

8.3. ...summary screen shows no data?
This could be due to the names of individuals not corresponding correctly between individuals in the fasta files and the demographic data text file. The demographic table needs to have a minimum of two columns, "Individual" and "Population". Make sure the names of the indivdiuals are the same as the fasta files. This could also be due to not importing a demographic data file. You need to import at least a list of the individuals and what population they belong to.

8.4. ...read data won't print?
Have you moved any folders? When first importing the data, a path is saved to the database, so if you move the sam/bam files, the program won't be able to find the folder. Try reimporting.

8.5. Problems with IMa2.
The IMa2InputFile.txt needs to look almost exactly like the one I've provided. Please let me know if you have specific problems with any of the output formats.

8.6. Problems with installation
First, check that you are using Python 2.7. Open a terminal and type "python". You should see something like this:

Python 2.7.1 (r271:86882M, Nov 30 2010, 10:35:34)
[GCC 4.2.1 (Apple Inc. build 5664)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>>

If your Python is version 2.5 or 2.6, try downloading Python 2.7 and reinstalling LOCINGS (section 1.2 above). If you have Python 2.7, type the following commands (after the >>>) to make sure all the packages installed correctly - if you get no feedback, it is installed correctly:
"import numpy"
"import Bio"
"import pysam"
"import pymongo"
"import simplejson"
"import cython"
(to exit the python prompt, type "exit()")

If you get an error like:
"Traceback (most recent call last):
File "<stdin>", line 1, in <module>
ImportError: No module named numpy"
then that package did not install correctly. You can go to the websites listed in section 1.3 above and install the problem package independently and rerun the installation steps in section 1.2 above.

If these suggestions do not work, please contact me with the error messages you're receiving and I'll try to help you get the program running. I understand there's something especially infuriating about buggy software, so please email me!

9. CONTACT
Please feel free to contact me about any issues you're having with LOCINGS or the dependent software. I'd be more than happy to do what I can.

Sarah Hird
sarah.hird@gmail.com

Appendix D.
Specimen Information For Birds Of Chapter 4.

Detailed sample information for birds used in Chapter 4, including the preparator's number, the LSUMNS bird collection number (B#), the proportion of skull ossification (Skull), adult/juvenile status (A/J), stomach contents (SC), broad dietary specialization, sampling locality (Loc) and the total number of high quality reads sequenced (Seqs).

Individual	Prep#	B#	Family	Genus	Species	Sex	Skull	A/J	SC	Diet	Loc	Seqs
C.card_01	DLD10182	61124	Cardinalidae	*Cardinalis*	*cardinalis*	m	1.00	JUV	nr	Herbivore	LA	12387
C.card_02	DLD10184	61126	Cardinalidae	*Cardinalis*	*cardinalis*	m	0.50	JUV	wm	Herbivore	LA	32433
C.card_03	DLD10186	61128	Cardinalidae	*Cardinalis*	*cardinalis*	f	1.00	AD	seeds	Herbivore	LA	13519
C.card_04.1	DLD10185	61227	Cardinalidae	*Cardinalis*	*cardinalis*	m	0.00	AD	wm	Herbivore	LA	91930
C.card_04.2	DLD10185	61227	Cardinalidae	*Cardinalis*	*cardinalis*	m	0.00	AD	wm	Herbivore	LA	86676
C.mexi_01	DLD10158	72199	Fringillidae	*Carpodacus*	*mexicanus*	m	0.00	JUV	seeds	Herbivore	CAL	90616
I.spur_01	DLD10199	81921	Icteridae	*Icterus*	*spurius*	m	1.00	AD	insects seeds	Omnivore	LA	65402
I.spur_02	DLD10200	81927	Icteridae	*Icterus*	*spurius*	m	1.00	AD	insects	Omnivore	LA	12982
M.ater_01	SWC8929	61074	Icteridae	*Molothrus*	*ater*	m	0.00	JUV	wm	Omnivore	LA	132574
M.ater_02	SWC8932	61077	Icteridae	*Molothrus*	*ater*	m	0.00	JUV	wm	Omnivore	LA	98047
M.ater_03	SWC8933	61078	Icteridae	*Molothrus*	*ater*	m	0.00	JUV	seeds	Omnivore	LA	73823
M.ater_04	SWC8934	61079	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	LA	43991
M.ater_05	SWC8935	61080	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	LA	43330
M.ater_06	JCC03	61109	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	60088
M.ater_07	DLD10176	61114	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	CAL	59577
M.ater_08	JCC04	61116	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	47239
M.ater_09	SWC8894	67666	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	17749
M.ater_10	SWC8896	67668	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	55077
M.ater_11	SWC8897	67669	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	93550
M.ater_12	SWC8901	67673	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	5112
M.ater_13	SWC8902	67674	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	19722
M.ater_14	SWC8903	67675	Icteridae	*Molothrus*	*ater*	m	1.00	AD	wm	Omnivore	LA	138875
M.ater_15	SWC8904	67676	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	32052
M.ater_16	DLD10168	67724	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	CAL	70309

Appendix D. Continued.

Individual	Prep#	B#	Family	Genus	Species	Sex	Skull	A/J	SC	Diet	Loc	Seqs
M.ater_17	DLD10169	67725	Icteridae	*Molothrus*	*ater*	f	0.10	JUV	wm	Omnivore	CAL	62475
M.ater_18	DLD10171	67727	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	CAL	89671
M.ater_19	DLD10172	67728	Icteridae	*Molothrus*	*ater*	m	0.00	JUV	wm	Omnivore	CAL	73390
M.ater_20	DLD10173	67729	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	CAL	33848
M.ater_21	SWC8911	67734	Icteridae	*Molothrus*	*ater*	m	1.00	AD	wm	Omnivore	LA	96809
M.ater_22	SWC8913	67736	Icteridae	*Molothrus*	*ater*	m	0.00	JUV	wm	Omnivore	LA	84006
M.ater_23	SWC8914	67737	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	68483
M.ater_24	SWC8915	67738	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	LA	56969
M.ater_25	SWC8919	67742	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	LA	72736
M.ater_26	SWC8920	67743	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	LA	59907
M.ater_27	DLD10159	72200	Icteridae	*Molothrus*	*ater*	f	1.00	AD	wm	Omnivore	CAL	43461
M.ater_28	DLD10160	72201	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	wm	Omnivore	CAL	68573
M.ater_29	DLD10162	72203	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	nr	Omnivore	CAL	52470
M.ater_30	DLD10163	72204	Icteridae	*Molothrus*	*ater*	m	0.00	JUV	nr	Omnivore	CAL	33010
M.ater_31	DLD10164	72205	Icteridae	*Molothrus*	*ater*	m	0.00	JUV	nr	Omnivore	CAL	70417
M.ater_32	DLD10166	72207	Icteridae	*Molothrus*	*ater*	f	0.00	JUV	nr	Omnivore	CAL	636535
P.cyan_01	DLD10195	81933	Cardinalidae	*Passerina*	*cyanea*	m	1.00	AD	insects seeds	Omnivore	LA	72221
P.caer_01	DLD10192	81919	Polioptilidae	*Polioptila*	*caerulea*	nr	0.00	JUV	insects	Carnivore	LA	55921
P.caer_02	DLD10193	81926	Polioptilidae	*Polioptila*	*caerulea*	f	nr	JUV	insects	Carnivore	LA	128044
P.citr_01	DLD10180	61122	Parulidae	*Prothonotaria*	*citrea*	f	1.00	AD	insects	Carnivore	LA	44351
P.citr_02	DLD10188	61130	Parulidae	*Prothonotaria*	*citrea*	m	0.00	JUV	insects fruit	Carnivore	LA	22906
T.ludo_01	DLD10187	61129	Troglodytidae	*Thryothorus*	*ludovicianus*	f	0.00	JUV	empty	Carnivore	LA	31090
T.ludo_02	DLD10191	81918	Troglodytidae	*Thryothorus*	*ludovicianus*	m	nr	AD	insects	Carnivore	LA	45235
V.gris_01	DLD10198	81920	Vireonidae	*Vireo*	*griseus*	m	1.00	AD	insects	Carnivore	LA	52393
W.citr_01	DLD10201	81905	Parulidae	*Wilsonia*	*citrina*	m	1.00	AD	insects	Carnivore	LA	48351

Appendix E.
Mammals and Insects Of Chapters 4 and 6.

Sample	Source	Order	Class	Specific Diet	Broad Diet	Seqs
Calosoma peregrinator	Colman	Coleoptera	insect	Predacious	Carnivore	248
Gonasida inferna	Colman	Coleoptera	insect	Omnivorous	Omnivore	230
Apis mellifera	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	271
Apis mellifera (hive)	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	267
Agapostemon virescens	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	273
Chalybion californicum	Colman	Hymenoptera	insect	Predacious	Carnivore	204
Coptotermes formosanus	Colman	Isoptera	insect	Xylophagous DeadWood	Herbivore	206
Colletes inaequalis	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	301
Calliopsis subalpinus	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	282
Caupolicana yarrowi	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	256
Diadasia opuntia	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	347
Hesperapis cockerelli	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	349
Halictus patellatus	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	305
Megachile odontostoma	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	338
Microcerotermes sp. M1	Colman	Isoptera	insect	Xylophagous DeadWood	Herbivore	217
Nasutitermes sp.	Colman	Isoptera	insect	Xylophagous DeadWood	Herbivore	1252
Philanthus gibbosus	Colman	Hymenoptera	insect	Predacious	Carnivore	360
Pieris rapae	Colman	Lepidoptera	insect	Herbivorous	Herbivore	1207
Paragia vespiformes	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	247
Reticulitermes speratus	Colman	Isoptera	insect	Xylophagous DeadWood	Herbivore	270
Rediviva saetigera	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	333
Xylocopa californica	Colman	Hymenoptera	insect	Pollenivorous	Herbivore	305
African Elephant	Ley	Proboscidae	mammal	Herbivore	Herbivore	991
Argali Sheep	Ley	Artiodactyla	mammal	Herbivore	Herbivore	540
Armadillo	Ley	Xenarthra	mammal	Carnivore	Carnivore	362
Asian Elephant	Ley	Proboscidae	mammal	Herbivore	Herbivore	412
Big Horn Sheep	Ley	Artiodactyla	mammal	Herbivore	Herbivore	618
Black Bear	Ley	Carnivora	mammal	Omnivore	Omnivore	372
Blackhanded Spider Monkey	Ley	Primates	mammal	Omnivore	Omnivore	272
Black Rhinoceros	Ley	Perissodactyla	mammal	Herbivore	Herbivore	287
Cheetah	Ley	Carnivora	mammal	Carnivore	Carnivore	267
Chimpanzee	Ley	Primates	mammal	Omnivore	Omnivore	290
Douc Langur	Ley	Primates	mammal	Herbivore	Herbivore	349
East Angolan Colubus	Ley	Primates	mammal	Herbivore	Herbivore	309
Eastern Black and White Colubus	Ley	Primates	mammal	Herbivore	Herbivore	204

Appendix E. Continued.

Sample	Source	Order	Class	Specific Diet	Broad Diet	Seqs
Echidna	Ley	Monotremata	mammal	Carnivore	Carnivore	393
Flying Fox	Ley	Chiroptera	mammal	Omnivore	Omnivore	228
Francois Langur	Ley	Primates	mammal	Herbivore	Herbivore	362
Geoffreys Marmoset	Ley	Primates	mammal	Omnivore	Omnivore	202
Giant Panda	Ley	Carnivora	mammal	Herbivore	Herbivore	564
Grevys Zebra	Ley	Perissodactyla	mammal	Herbivore	Herbivore	207
Hamadryas Baboon	Ley	Primates	mammal	Omnivore	Omnivore	367
Hedgehog	Ley	Insectivora	mammal	Carnivore	Carnivore	211
Horse	Ley	Perissodactyla	mammal	Herbivore	Herbivore	509
Lion	Ley	Carnivora	mammal	Carnivore	Carnivore	409
Molerat	Ley	Rodentia	mammal	Herbivore	Herbivore	318
Mongoose Lemur	Ley	Primates	mammal	Omnivore	Omnivore	281
Okapi	Ley	Artiodactyla	mammal	Herbivore	Herbivore	391
Orangutan	Ley	Primates	mammal	Herbivore	Herbivore	497
Polar Bear	Ley	Carnivora	mammal	Carnivore	Carnivore	485
Red Kangaroo	Ley	Diprotodontia	mammal	Herbivore	Herbivore	273
Red Panda	Ley	Carnivora	mammal	Herbivore	Herbivore	1144
Sebas Short Tailed Bat	Ley	Chiroptera	mammal	Omnivore	Omnivore	268
Spekes Gazelle	Ley	Artiodactyla	mammal	Herbivore	Herbivore	454
Spotted Hyena	Ley	Carnivora	mammal	Carnivore	Carnivore	257
Squirrel	Ley	Rodentia	mammal	Omnivore	Omnivore	222
Transcaspian Urial Sheep	Ley	Artiodactyla	mammal	Herbivore	Herbivore	301
Western Lowland Gorilla	Ley	Primates	mammal	Herbivore	Herbivore	468
Whitefaced Saki	Ley	Primates	mammal	Omnivore	Omnivore	416
Zebra	Ley	Perissodactyla	mammal	Herbivore	Herbivore	229

Appendix F.
Specimen Information For Birds Of Chapter 5.

Detailed information about each sample used in Chapter 5, including taxonomic information, dietary assignments, habitat, foraging stratum (FS), sampling locality (L), elevation (Elev), percent of skull ossification (Age), bacterial phyla identified (BRP), the percentage of individuals with fewer phylotypes identified (BRS) and LSUMNS bird tissue collection number (B#).

Genus species (ID)	Order	Family	Diet Specific	Diet Br	Diet BO	Habitat	FS	L	Elev (m)	Age	Stomach Contents	BRP	BRS (%)	Sex	B#
Amazilia tzacatl	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U_C	G	170	0	nr	13	100	F	71825
Florisuga mellivora 1	Apodiformes	Trochilidae	nectar insect	o	NE	FT SC GR	M_C	C	260	10	insect	13	20	F	71967
Florisuga mellivora 2	Apodiformes	Trochilidae	nectar insect	o	NE	FT SC GR	M_C	L	415	70	insect pollen	13	80	M	73941
Florisuga mellivora 3	Apodiformes	Trochilidae	nectar insect	o	NE	FT SC GR	M_C	G	170	0	nr	12	60	M	71880
Florisuga mellivora 4	Apodiformes	Trochilidae	nectar insect	o	NE	FT SC GR	M_C	C	260	0	e	12	60	F	71964
Phaethornis longirostris 1	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U	C	65	10	e	15	60	F	71984
Phaethornis longirostris 2	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U	B	110	3	nr	14	60	M	71935
Phaethornis longirostris 3	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U	B	110	0	e	14	60	F	71928
Thalurania colombica 1	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U_M	C	65	0	nr	8	20	M	71982
Thalurania colombica 2	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U_M	C	65	10	insect	15	60	nr	71985
Threnetes ruckeri 1	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U	G	170	20	e	10	20	F	71857
Threnetes ruckeri 2	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U	B	110	5	nr	12	80	F	71920
Threnetes ruckeri 3	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U	G	170	0	nr	12	40	F	71869

Appendix F. Continued.

Genus species (ID)	Order	Family	Diet Specific	Diet Br	Diet BO	Habitat	FS	L	Elev (m)	Age	Stomach Contents	BRP	BRS (%)	Sex	B#
Threnetes ruckeri 4	Apodiformes	Trochilidae	nectar	h	NE	FT SC GR	U	G	170	20	e	9	40	M	71853
Nyctidromus albicollis 1.1	Caprimulgiformes	Caprimulgidae	insect	c	IN	WO	T	I	430	100	insect	12	20	nr	71999
Nyctidromus albicollis 1.2	Caprimulgiformes	Caprimulgidae	insect	c	IN	WO	T	I	430	100	insect	11	40	nr	71999
Geotrygon montana	Columbiformes	Columbidae	seed	h	FR	WO	T	C	65	95	plant	14	40	F	71937
Baryphthengus martii	Coraciiformes	Momotidae	arthropod vert	c	IN	FT	U_M	G	170	100	nr	12	60	F	71827
Piaya cayana	Cuculiformes	Cuculidae	insect	c	IN	WO	C	I	430	100	insect	10	20	M	71998
Cyanocompsa cyanoides 1 1	Passeriformes	Cardinalidae	seed	h	FR	WO	U	G	170	50	e	14	60	F	71872
Cyanocompsa cyanoides 1 2	Passeriformes	Cardinalidae	seed	h	FR	WO	U	G	170	50	e	12	40	F	71872
Cyanocompsa cyanoides 2	Passeriformes	Cardinalidae	seed	h	FR	WO	U	K	325	100	nr	13	100	nr	74234
Cyanocompsa cyanoides 3	Passeriformes	Cardinalidae	seed	h	FR	WO	U	C	260	100	plant	12	40	F	71965
Cyanocompsa cyanoides 4	Passeriformes	Cardinalidae	seed	h	FR	WO	U	K	325	100	seeds	13	100	M	74232
Cyanocompsa cyanoides 5	Passeriformes	Cardinalidae	seed	h	FR	WO	U	L	250	100	plant	17	100	M	73992
Cyanocompsa cyanoides 6	Passeriformes	Cardinalidae	seed	h	FR	WO	U	K	325	5	nr	15	80	M	74155
Cyanocompsa cyanoides 7	Passeriformes	Cardinalidae	seed	h	FR	WO	U	C	65	100	e	11	40	M	71944
Habia atrimaxillaris 1	Passeriformes	Cardinalidae	insect frug	o	FR	WO	U	B	110	100	insect	13	100	F	71918
Habia atrimaxillaris 2	Passeriformes	Cardinalidae	insect frug	o	FR	WO	U	C	65	100	nr	12	40	F	71943
Habia fuscicauda 1	Passeriformes	Cardinalidae	insect frug	o	FR	WO	U	G	170	100	fruit insect	11	40	F	71835
Habia fuscicauda 2	Passeriformes	Cardinalidae	insect frug	o	FR	WO	U	G	170	100	e	10	100	F	71851

Appendix F. Continued.

Genus species (ID)	Order	Family	Diet Specific	Diet Br	Diet BO	Habitat	FS	L	Elev (m)	Age	Stomach Contents	BRP	BRS (%)	Sex	B#
Habia fuscicauda 3	Passeriformes	Cardinalidae	insect frug	o	FR	WO	U	G	170	5	nr	14	100	F	71832
Habia fuscicauda 4	Passeriformes	Cardinalidae	insect frug	o	FR	WO	U	G	170	100	insect	12	40	M	71839
Arremon aurantiirostris 1	Passeriformes	Emberizidae	generalist	o	FR	WO	T	D	200	100	seeds insects	15	100	F	71811
Arremon aurantiirostris 2	Passeriformes	Emberizidae	generalist	o	FR	WO	T	G	170	5	insect	11	20	M	71858
Arremonops conirostris	Passeriformes	Emberizidae	generalist	o	FR	WO	T_U	C	65	5	seeds	13	60	F	71995
Formicarius analis 1	Passeriformes	Formicariidae	arthropod	c	IN	FT	T	D	200	100	insect	8	20	M	71809
Formicarius analis 2	Passeriformes	Formicariidae	arthropod	c	IN	FT	T	B	110	50	insect	13	80	M	71933
Automolus ochrolaemus	Passeriformes	Furnariidae	insect	c	IN	FT	U	A	75	100	insect	14	20	nr	72780
Dendrocincla fuliginosa	Passeriformes	Furnariidae	arthropod	c	IN	FT	U_M	G	170	100	e	14	100	M	71846
Glyphorhynchus spirurus 1	Passeriformes	Furnariidae	insect	c	IN	FT	U_M	I	430	100	nr	9	20	M	72001
Glyphorhynchus spirurus 2	Passeriformes	Furnariidae	insect	c	IN	FT	U_M	H		100	nr	14	100	F	71820
Xiphorhynchus susurrans 1	Passeriformes	Furnariidae	arthropod	c	IN	FT	U_C	C	65	90	insect	11	20	M	71993
Xiphorhynchus susurrans 2	Passeriformes	Furnariidae	arthropod	c	IN	FT	U_C	C	65	100	e	13	20	F	71941
Xiphorhynchus susurrans 3	Passeriformes	Furnariidae	arthropod	c	IN	FT	U_C	G	170	100	e	10	20	M	71861
Cacicus uropygialis 1	Passeriformes	Icteridae	generalist	o	FR	WO	C	C	260	100	seeds	11	40	M	71968
Cacicus uropygialis 2	Passeriformes	Icteridae	generalist	o	FR	WO	C	F	1050	100	seeds insects	13	40	nr	72046
Cacicus uropygialis 3	Passeriformes	Icteridae	generalist	o	FR	WO	C	C	260	10	fruit	12	60	M	71969

Appendix F. Continued.

Genus species (ID)	Order	Family	Diet Specific	Diet Br	Diet BO	Habitat	FS	L	Elev (m)	Age	Stomach Contents	BRP	BRS (%)	Sex	B#
Saltator maximus	Passeriformes	IncertaeSedis	generalist	o	FR	WO	M_C	G	170	10	nr	12	80	nr	71830
Myiothlypis fulvicauda	Passeriformes	Parulidae	arthropod	c	FR	WO	T	E	400	15	insect	15	80	M	72059
Manacus candei 1	Passeriformes	Pipridae	frug	h	IN	FT GR	U	G	170	100	fruit	14	40	F	71826
Manacus candei 2	Passeriformes	Pipridae	frug	h	IN	FT GR	U	G	170	100	e	11	20	F	71833
Manacus candei 3	Passeriformes	Pipridae	frug	h	IN	FT GR	U	G	170	25	e	14	80	M	71849
Manacus candei 4	Passeriformes	Pipridae	frug	h	IN	FT GR	U	G	170	100	e	13	60	F	71836
Manacus candei 5	Passeriformes	Pipridae	frug	h	IN	FT GR	U	G	170	100	nr	12	60	F	71823
Manacus candei 6	Passeriformes	Pipridae	frug	h	IN	FT GR	U	G	170	100	fruit	12	80	F	71866
Manacus aurantiacus	Passeriformes	Pipridae	frug	h	IN	FT GR	U	C	65	nr	plant	14	80	F	71991
Pipra mentalis 1	Passeriformes	Pipridae	frug	h	IN	FT GR	U_M	D	200	25	nr	13	80	M	71807
Pipra mentalis 2	Passeriformes	Pipridae	frug	h	IN	FT GR	U_M	G	170	100	e	14	80	F	71892
Pipra mentalis 3	Passeriformes	Pipridae	frug	h	IN	FT GR	U_M	C	65	50	e	14	60	nr	71953
Pipra mentalis 4	Passeriformes	Pipridae	frug	h	IN	FT GR	U_M	G	170	100	e	13	40	M	71881
Pipra mentalis 5	Passeriformes	Pipridae	frug	h	IN	FT GR	U_M	G	170	100	e	12	20	F	71875
Pipra mentalis 6	Passeriformes	Pipridae	frug	h	IN	FT GR	U_M	B	110	100	e	11	60	M	71930
Pipra mentalis 7	Passeriformes	Pipridae	frug	h	IN	FT GR	U_M	C	65	100	e	9	20	M	71940
Cymbilaimus lineatus	Passeriformes	Thamnophilidae	insect	c	IN	FT	C	J	80	100	insect	12	100	M	72020

Appendix F. Continued.

Genus species (ID)	Order	Family	Diet Specific	Diet Br	Diet BO	Habitat	FS	L	Elev (m)	Age	Stomach Contents	BRP	BRS (%)	Sex	B#
Gymnopithys leucaspis 1	Passeriformes	Thamnophilidae	insect	c	IN	FT	U	C	65	100	e	12	60	F	71957
Gymnopithys leucaspis 2	Passeriformes	Thamnophilidae	insect	c	IN	FT	U	G	170	100	e	14	100	F	71860
Gymnopithys leucaspis 3	Passeriformes	Thamnophilidae	insect	c	IN	FT	U	C	65	100	e	14	40	M	71955
Hylophylax naevioides	Passeriformes	Thamnophilidae	insect	c	IN	FT	U	G	170	100	e	15	80	F	71854
Microrhopias quixensis 1	Passeriformes	Thamnophilidae	insect	c	IN	FT	M	F	1050	10	insect	17	100	nr	72125
Microrhopias quixensis 2	Passeriformes	Thamnophilidae	insect	c	IN	FT	M	C	260	100	insect	13	40	M	71966
Myrmeciza exsul 2	Passeriformes	Thamnophilidae	arthropod	c	IN	FT	U	J	80	50	insect	17	60	F	72021
Myrmeciza exsul 1	Passeriformes	Thamnophilidae	arthropod	c	IN	FT	U	G	170	100	e	15	80	M	71873
Chlorophanes spiza	Passeriformes	Thraupidae	frug nectar insect	o	FR	WO	C	C	260	90	nr	13	60	F	71963
Oryzoborus funereus	Passeriformes	Thraupidae	seed	h	FR	WO	U_M	G	170	10	seeds plant	13	100	M	71856
Ramphocelus costaricensis	Passeriformes	Thraupidae	insect frug	o	FR	WO	U_C	C	65	10	e	12	20	M	72016
Ramphocelus passerinii 1	Passeriformes	Thraupidae	frug insect	o	FR	WO	U_C	G	170	100	e	11	40	F	71844
Ramphocelus passerinii 2	Passeriformes	Thraupidae	frug insect	o	FR	WO	U_C	G	170	5	e	11	60	M	71886
Ramphocelus passerinii 3	Passeriformes	Thraupidae	frug insect	o	FR	WO	U_C	G	170	100	insect	12	20	M	71852
Sporophila corvina	Passeriformes	Thraupidae	seed	h	FR	WO	U_M	G	170	5	seeds	14	80	nr	71862
Tachyphonus luctuosus	Passeriformes	Thraupidae	insect frug	o	FR	WO	M_C	I	430	100	insect	17	100	F	72011
Tangara larvata 1	Passeriformes	Thraupidae	insect frug	o	FR	WO	C	I	430	50	seeds plant	15	80	M	71997

Appendix F. Continued.

Genus species (ID)	Order	Family	Diet Specific	Diet Br	Diet BO	Habitat	FS	L	Elev (m)	Age	Stomach Contents	BRP	BRS (%)	Sex	B#
Tangara larvata 2	Passeriformes	Thraupidae	insect frug	o	FR	WO	C	G	170	3	e	12	20	nr	71870
Tangara gyrola	Passeriformes	Thraupidae	insect frug	o	FR	WO	C	C	260	100	seeds plant	11	60	F	71973
Volatinia jacarina	Passeriformes	Thraupidae	seed	h	FR	WO	T_U	G	170	75	e	15	80	F	71841
Thraupis episcopus	Passeriformes	Thraupis	insect frug	o	FR	WO	C	G	170	3	e	13	40	M	71910
Tityra inquisitor	Passeriformes	Tityridae	insect frug	o	IN	FT GR	C	C	65	100	plant	17	100	F	71954
Cantorchilus nigricapillus	Passeriformes	Troglodytidae	arthropod	c	IN	WO	U	I	260	100	insect	9	20	M	72007
Henicorhina leucosticta 1	Passeriformes	Troglodytidae	arthropod	c	IN	WO	U	I	260	100	insect	15	100	M	72008
Henicorhina leucosticta 2	Passeriformes	Troglodytidae	arthropod	c	IN	WO	U	G	170	100	insect	12	80	F	71837
Turdus grayi	Passeriformes	Turdidae	generalist	o	IN	ALL	T_M	G	170	100	nr	14	60	M	71834
Attila spadiceus 1.1	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	M_C	A	75	100	insect	15	100	M	72081
Attila spadiceus 1.2	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	M_C	A	75	100	insect	13	100	M	72081
Elaenia flavogaster	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	C	G	170	50	e	13	40	F	71877
Mionectes oleagineus 1	Passeriformes	Tyrannidae	frug	h	IN	FT GR	U_C	K	325	20	seeds	13	80	F	74190
Mionectes oleagineus 2	Passeriformes	Tyrannidae	frug	h	IN	FT GR	U_C	L	250	10	e	15	80	F	74000
Mionectes oleagineus 3	Passeriformes	Tyrannidae	frug	h	IN	FT GR	U_C	C	65	25	seeds	12	80	M	71850
Mionectes oleagineus 4	Passeriformes	Tyrannidae	frug	h	IN	FT GR	U_C	B	110	50	e	14	100	M	71932
Mionectes oleagineus 5	Passeriformes	Tyrannidae	frug	h	IN	FT GR	U_C	G	170	25	e	14	40	nr	71867
Myiarchus tuberculifer	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	M_C	C	65	100	e	15	100	M	72089

Appendix F. Continued.

Genus species (ID)	Order	Family	Diet Specific	Diet Br	Diet BO	Habitat	FS	L	Elev (m)	Age	Stomach Contents	BRP	BRS (%)	Sex	B#
Myiozetetes granadensis	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	C	G	170	100	insect plant	11	20	F	71891
Myiozetetes similis 1	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	M_C	C	65	75	e	13	80	M	72015
Myiozetetes similis 2	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	M_C	C	65	75	insect	16	100	F	72014
Myiozetetes similis 3	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	M_C	C	65	100	e	13	20	M	72013
Onychorhynchus coronatus	Passeriformes	Tyrannidae	insect	c	IN	FT GR	M	G	170	100	e	13	60	nr	71871
Platyrinchus coronatus	Passeriformes	Tyrannidae	arthropod	c	IN	FT GR	U_M	B	110	100	insect	16	100	M	71923
Tolmomyias sulphurescens	Passeriformes	Tyrannidae	insect frug	o	IN	FT GR	C	G	170	50	e	12	80	F	71890
Hylophilus flavipes	Passeriformes	Vireonidae	insect	c	IN	FT	M_C	A	75	100	fruit	16	100	M	72075
Galbula ruficauda 1	Piciformes	Galbulidae	insect	c	IN	FT	M	G	170	100	e	9	20	M	71831
Galbula ruficauda 2	Piciformes	Galbulidae	insect	c	IN	FT	M	G	170	100	nr	13	60	F	71828
Melanerpes pucherani	Piciformes	Picidae	insect frug	o	IN	WO	C	G	170	100	fruit insect seeds	13	40	M	71909
Pteroglossus torquatus	Piciformes	Ramphastidae	frug insect	o	FR	FT	C	F	1050	100	fruit	14	100	nr	72054
Trogon massena	Trogoniformes	Trogonidae	generalist	o	OM	FT	M_C	I	430	100	fruit insect	11	40	F	72010
Trogon rufus 1	Trogoniformes	Trogonidae	insect	c	OM	FT	U_M	B	110	25	seeds	14	60	F	71929
Trogon rufus 2.1	Trogoniformes	Trogonidae	insect	c	OM	FT	U_M	I	260	100	fruit insect	15	80	M	72006
Trogon rufus 2.2	Trogoniformes	Trogonidae	insect	c	OM	FT	U_M	I	260	100	fruit insect	14	40	M	72006

Diet Br: h = mostly plant, o = plant and animal, c = mostly animal; Diet BO: FR = frugivore, IN = invertebrates, OM = omnivore, NE = nectar
Habitat: FT = forest, GR = grassland/steppe/savanah, SC = scrub, WO = woodland; Foraging Strata (FS): T = terrestrial, U = understory, M = midcanopy, C = canopy
Locality: See Figure 5.1.; Stomach contents: e = empty, nr = not recorded; Sex: M = male, F = female, nr = not recorded

Appendix G.
Permissions from Cambridge University Press

PERMISSION INVOICE

Inv. # P03B 22579

February 20, 2013

Sarah M Hird
119 Foster Hall
Museum of Natural Science
Louisiana State University
Baton Rouge, LA 70803

CAMBRIDGE UNIVERSITY PRESS
32 Avenue of the Americas
New York, NY 10013-2473, USA
www.cambridge.org
Telephone 212 924 3900
Fax 212 691 3239

REFERENCE

ISBN:	HB 9780521444187 PB Other
Author:	C. Edward Stevens and Ian D. Hume
Title:	COMPARATIVE PHYSIOLOGY OF THE VERTEBRATE DIGESTIVE SYSTEM, 2ND EDITION
Selection/pp.:	Figure 3.13, pg. 42 and Figure 3.14, pg. 43
Additional:	Copyright © 1995 Cambridge University Press.

USE

Reprint Title:	NOVEL COMPUTATIONAL TOOLS AND UTILIZATION OF THE GUT MICROBIOTA FOR PHYLOGEOGRAPHIC INFERENCE
Publisher:	Louisiana State University
Format:	dissertation / thesis
Quantity (Limit*):	120
Avail. Date:	2013

RIGHTS/ACKNOWLEDGEMENT

Permission is granted for nonexclusive rights throughout the World in the English language for interior text editorial use in the format described above only, including non-profit editions for the blind and handicapped. Please fully acknowledge our material and indicate the copyright notice as it appears in our publication, followed by the phrase "Reprinted with the permission of Cambridge University Press."
All requests from third parties to reproduce this material must be forwarded to Cambridge University Press.

FEES/RESTRICTIONS

$0.00

*This permission is restricted to the indicated format and quantity; for additional use, you must reapply for permission. This permission does not allow reprinting of any material copyrighted by or credited in our publication to another source; Cambridge disclaims all liability in connection with the use of such material without proper consent.
A COPY OF THIS INVOICE MUST ACCOMPANY PAYMENT. Payment is due upon receipt of invoice. Terms: Net 60 days. Make check payable to Cambridge University Press, Attn: Rights and Permissions. (Fed. I.D. #: 13-1599108.)

This permission does not supersede permission that may be required from the original source indicated in our publication.
This permission requires that you send zero (0) copies of your publication directly to our author and zero (0) copy of your publication to this office upon availability.

Authorization: Adam Hirschberg

Adam Hirschberg
Rights and Permissions Associate
ahirschberg@cambridge.org

Appendix H.
Permissions For Portions of Chapter 1

ELSEVIER LICENSE
TERMS AND CONDITIONS

Feb 13, 2013

This is a License Agreement between Sarah M Hird ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Sarah M Hird
Customer address	119 Foster Hall Baton Rouge, LA 70803
License number	3087180231059
License date	Feb 13, 2013
Licensed content publisher	Elsevier
Licensed content publication	Molecular Phylogenetics and Evolution
Licensed content title	Applications of next-generation sequencing to phylogeography and phylogenetics
Licensed content author	John E. McCormack,Sarah M. Hird,Amanda J. Zellmer,Bryan C. Carstens,Robb T. Brumfield
Licensed content date	February 2013
Licensed content volume number	66
Licensed content issue number	2
Number of pages	13
Start Page	526
End Page	538
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	excerpt
Number of excerpts	2
Format	both print and electronic
Are you the author of this	Yes

Appendix I.
Permissions For Chapter 2

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Dec 04, 2012

This is a License Agreement between Sarah M Hird ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3042021035251
License date	Dec 04, 2012
Licensed content publisher	John Wiley and Sons
Licensed content publication	Molecular Ecology Resources
Book title	
Licensed content author	SARAH M. HIRD,ROBB T. BRUMFIELD,BRYAN C. CARSTENS
Licensed content date	Mar 24, 2011
Start page	743
End page	748
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print
Portion	Full article
Will you be translating?	No
Order reference number	
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at http://myaccount.copyright.com)

Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

Appendix J.
Permissions For Chapter 3

Public Library of Science
1160 Battery St, Suite 100
San Francisco, CA 94111

TIN: 68-0492065 Tax Exempted

Invoice

Email: authorbilling@plos.org
Phone: (415)568-4502
Secure Fax: (415)568-3459

Bill To

HIRD, SARAH
Louisiana State University
Biological Sciences and the Museum of Natural Science
Museum of Natural Science
119 Foster Hall
Baton Rouge, LA 70803

Date	09/10/12
Invoice #	PAB54158
Payment Terms	Due on Receipt
PO #	

Article Number	**Title of Manuscript**			**Journal Name**	
PONE-D-12-21382	lociNGS: a lightweight alternative for assessing suitability of next-generation loci for evolutionar			PLoS ONE	
Purpose of Invoice	**Publication fee for the above referenced article**				
Total Publication Fee	500.00	USD	**Balance Due**	500.00	USD

The publication charge is a flat fee per article published.
The amount includes the cost of peer review, journal production, online hosting & archiving.

Payment Methods:

Please include the invoice number with all payments
Check must be in U.S. Dollars & drawn on a U.S. bank
For EFT or ACH please use 121000358 (U.S. Only)

By Check:

Make Payable to: PLoS or Public Library of Science
Mailing Address: Public Library of Science, 1160 Battery St, Suite 100, San Francisco, CA 94111

By Wire:

Name of Bank: Bank of America
Bank Address: 2129 Shattuck Avenue, Berkeley, CA 94704 United States

Beneficiary: Public Library of Science
Bank Account #: 0175171692
SWIFT code: BOFAUS3N
Wire Routing #: 026009593

For wire payments please transfer the amount indicated above in U.S. Dollars & include any applicable wire transfer fees.
Please ensure that the invoice number is quoted on the wire advice.

By Credit Card:

Please click on the secure link below and it will take you to our PLoS payment and FAQ's page

http://www.plos.org/publish/pricing-policy/pay-invoice/

If the link does not work, please copy and paste it into your web browser.

AS A NON-PROFIT, ALL OF OUR REVENUE IS USED TO FURTHER OUR OPEN ACCESS MISSION.
IF PAYING BY CREDIT CARD, WE KINDLY REQUEST THAT YOU INCLUDE AN ADDITIONAL $25
TO HELP DEFRAY OUR BANK CHARGES. (OPTIONAL)

Vita

Sarah Michelle Hird was born in Rochester, Minnesota in 1983. She moved to Austin, Texas at an early age and moved again to Boise, Idaho at the age of fourteen. She attended Eagle High School and was Valedictorian of her class, a National Merit Scholar and President of the Eagle High chapter of the National Honor Society. She also worked part-time for three years at a wonderful Italian restaurant, daVinci's, and was the school mascot (a mustang) for a year.

Sarah attended the University of Idaho as a biology major, beginning in August 2001. She worked as an undergraduate research assistant for Dr. Kari Segraves in Dr. Jack Sullivan's lab and for Dr. David Althoff in Dr. Olle Pellmyr's lab. She was also a biology teaching assistant for non-majors. She continued at the University of Idaho for a Master of Science with Dr. Sullivan in 2005. Another Master's student started that same fall, Noah Mattoon Reid, and he and Sarah became good friends quickly. Not quite as quickly, they became sweethearts.

Noah and Sarah moved to Baton Rouge in August of 2008, so Noah could begin a PhD at LSU. After a year off, Sarah returned to Academia in August 2009, beginning a PhD with Dr. Robb Brumfield and Dr. Bryan Carstens, a decision she will always be grateful for. Sarah and Noah got married in October 2010 and will both be starting post-doctoral positions at the University of California, Davis, in Fall 2013. They plan to live happily ever after. Geaux Tigers.